OVERCOMING DATA CHALLENGES

IN MACHINE TRANSLATION

by

Huda Khayrallah

A dissertation submitted to The Johns Hopkins University

in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

July, 2021

Abstract

Data-driven machine translation paradigms—which use machine learning to create translation models that can automatically translate from one language to another—have the potential to enable seamless communication across language barriers, and improve global information access. For this to become a reality, machine translation must be available for all languages and styles of text. However, the translation quality of these models is sensitive to the quality and quantity of the data the models are trained on. In this dissertation we address and analyze challenges arising from this sensitivity; we present methods that improve translation quality in difficult data settings, and analyze the effect of data quality on machine translation quality.

Machine translation models are typically trained on parallel corpora, but limited quantities of such data are available for most language pairs, leading to a low resource problem. We present a method for transfer learning from a paraphraser to overcome data sparsity in low resource settings. Even when training data is available in the desired language pair, it is frequently of a different style or genre than we would like to translate—leading to a domain mismatch. We present a method for improving

domain adaptation translation quality.

A seemingly obvious approach when faced with a lack of data is to acquire more data. However, it is not always feasible to produce additional human translations. In such a case, an option may be to crawl the web for additional training data. However, as we demonstrate, such data can be very noisy and harm machine translation quality. Our analysis motivated subsequent work on data filtering and cleaning by the broader community.

The contributions in this dissertation not only improve translation quality in difficult data settings, but also serve as a reminder to carefully consider the impact of the data when training machine learning models.

Primary Reader and Advisor: Philipp Koehn

Secondary Readers: Kevin Duh & Matt Post

Acknowledgments

I am incredibly grateful for all the support I have received during my PhD, and leading up to it. I recognize these acknowledgments will be both too long, and not enough.

Committee

Thank you to my advisor, Philipp Koehn, for the freedom and faith in me. Philipp has an incredibly open view on research collaboration, which created a PhD in which I had the opportunity to learn from many others as well. I appreciated learning not only research from him, but also his philosophies on how to communicate that research and present it. Thank you to my other committee members, Kevin Duh and Matt Post. Kevin for the guidance & support, and letting me be your occasional honorary student. Kevin has a knack for seeing future research questions in every paper, and asking interesting research questions—two qualities I strive to emulate. Matt for the enthusiasm and excitement, and always being ready to dive into something new. Matt also has an admirable commitment to professional service (e.g. ACL Anthology and

SacreBleu). His work has not only enabled my own research, but also drives the community forward.

JHU Faculty

Thank you to Ben Van Durme for the mentorship and for looking out for me; Raman Arora for the early project supervision; and Paul McNamee for jovial kindness. Thanks to João Sedoc for the openness to work with someone naive about dialog, which led to a fun and productive collaboration in COVID times.

NLP Researchers

I am grateful to have received additional NLP advice and mentoring from Amittai Axelrod, Chris Callison-Burch, Marine Carpuat, Hal Daumé III, Frank Ferraro, Nizar Habash, and Graham Neubig.

JHU Staff

I am indebted to Ruth Scally, Yamese Diggs, Carl Pupa, Jaime Tebas-Pueyo, Steve Hart, Jennifer Linton, and Lauren Bigham for keeping CLSP and its compute cluster running, and collectively saving me countless days (if not weeks) of time. Thank you as well to Kim Franklin and Zack Burwell for CS related assistance.

ACKNOWLEDGMENTS

Collaborators

I had the opportunity to co-author with a variety of people, who each taught me something during our collaborations. Thank you to Philipp Koehn, Kevin Duh, Matt Post, Brian Thompson, Gaurav Kumar, Jeremy Gwinnup, Biman Gujral, Xuan Zhang, Rebecca Knowles, Shuoyang Ding, Paul McNamee, Kenton Murray, João Sedoc, David Yarowsky, Chris Kirov, Arya McCarthy, Winston Wu, Tongfei Chen, Tim Anderson, Steven Shearing, Sheng Zhang, Sean Trott, Ryan Culkin, Ryan Cotterell, Rebecca Marvin, Raman Arora, Patrick Xia, Mikel Forcada, Marine Carpuat, Marianna Martindale, Kenneth Heafield, Jerome Feldman, Jacob Bremerman, Hainan Xu, Ekaterina Vylomova, Edward Hu, Dee Ann Reisinger, Colleen Lewis, Benjamin Van Durme, Antonios Anastasopoulos, Amy Tsai, and Adrian Benton.

JHU CLSP

One of the benefits of JHU's CLSP is the large community, who I had the opportunity to learn from, and many of whom became friends in addition to colleagues.

Thanks to Rachel Rudinger, for the advice, friendship, shared meals (and time spent deciding on those meals).

Rebecca Knowles was the first of Philipp's JHU students to defend and it was always comforting to have someone a few steps ahead to ask for guidance. Shuoyang Ding was the first fully-Philipp-advised fully-JHU student, and provided similar perspective of someone one step ahead. Gaurav Kumar could provide background,

experience, and perspective on just about any topic or problem I could think of, and Adi Renduchintala could get excited about new directions from just about any topic or problem I could think of. Both also provided much need comic relief. Thanks to Brian Thompson for the sustained collaboration, practical perspectives, and being the first to defend in COVID times (and sharing lessons learned). Liz Salesky both had plenty of enthusiasm to share, and was a calming sounding board.

Thanks to Rachel Wicks for the excitement and fresh perspectives, and to Carlos Aguirre, for wanting and working to make CLSP & CS a better place. Elliot Schumacher was always the refreshingly reasonable one, thanks for being willing to advocate to spread that reason.

Thanks to Craig Harman and Matthew Wiesner for the camaraderie during LORELEI evaluations and PI meetings. Jonathan Jones was a fantastic desk neighbor and provided an ideal mix of: outside perspectives on my work, encouragement, new recipes, and adorable stories. Thank you to Keisuke Sakaguchi for being such a pleasant office mate, and for the original version of this thesis template. Thanks to Zach Wood-Doughty for often lending a late-night ear, and Patrick Xia for making our office a community (though shared meals and animal live-streams).

Thanks to Biman Gujral for the collaboration: both on research, and on baked goods. Adam Poliak provided many laughs, and an inside connection to all things JHU. Thank you to Chandler May for the quiet kindness. Thanks to Winston Wu for the Great Wall trips, and Hainan Xu & Jonathan Jones for the company on them.

ACKNOWLEDGMENTS

Patrick Xia, Marc Marone, Nathaniel Weir and Xuan Zhang provided valuable detailed feedback at key times, and Sabrina Mielke provided (principled, and less principled) LaTeX help.

Special thanks to 'my 1st years'—both the cohort who started in 2016 (the year after me), and those who started in 2019 (my last in-person fall semester)—who had contagious enthusiasm and comradery. Thanks for brightening up the office, and letting some of us 'old ones' crash occasionally.

Thank you to students I TAed, and those whose research projects I mentored (Steven Shearing, Jacob Bremerman, Lisa Zhu, Shuhao Lai, Ishita Tripathi, and Steven Tan)—your enthusiasm was my motivation.

MT Communities

Thanks to the (broader) MT group at JHU for the community, discussions, and (*lots of*) paper proofreading & practice talks—Shuoyang Ding, Kevin Duh, Mitchell Gordon, Biman Gujral, Jeremy Gwinnup, Rebecca Knowles, Philipp Koehn, Gaurav Kumar, Xutai Ma, Kelly Marchisio, Becky Marvin, Kenton Murray, Matt Post, Adi Renduchintala, Elizabeth Salesky, Pamela Shapiro, Brian Thompson, Rachel Wicks, Winston Wu, and Xuan Zhang.

My colleagues from my internship at Lilt—including Gabriel Bretschner, Juan Cabello, John DeNero, Spence Green, Ashwin Purohit, Patrick Simianer, Joern Wuebker, and Thomas Zenkel—provided me with practical MT perspectives, new

research ideas, and a welcome change of scenery.

The US MT community has a lovely event—MT marathon in the Americas–which bring together researchers from across the county (and a few from abroad). Thanks to the organizers—Timothy Anderson, Marine Carpuat, David Chiang, Christian Federmann, Jeremy Gwinnup, Philipp Koehn, Graham Neubig, and Lane Schwartz—for making it happen.

JHU Communities

Thank you to Kate Fischl for the initiative to co-found GRACE, and for the continued friendship. Thanks to Alycen Wiacek for being a friend and (early-morning) sounding board, and thanks as well to Ayushi Sinha, Arlene Chiu, Michelle Graham, and Kelly Marchisio for the community. Thank you to Yasamin Nazari for striving to make the CS department a community, and a better one. Thanks to Rachel Sherman and Benj Shapiro for the ongoing grad school check-ins (always over very tasty food).

Though it was short-lived, grad-student book club provided a welcome excuse to get back in to reading and hang out with Jeff Craley, Jonathan Jones, Jane Lutken, Sadhwi Srinivas, and Zach Wood-Doughty (among others).

During my time in Baltimore, I had a rotating collection of housemates, who provided much needed balance. In particular, thank you to Lauren Chambers, Ty Pan, and Isa Trejo-Zambrano. Lauren & Ty for being sounding boards and consistently reminding I was not ridiculous (and providing much needed breaks). And Isa for

balancing my ridiculousness as we navigated the pandemic while I job hunted and wrapped up my PhD.

Berkeley

Thank you to Colleen Lewis and Floraine Berthouzoz for starting CS KickStart. Colleen, thank you for the first step into research, and for the continued mentorship over the last 10 years. Floraine, thank you for recognizing that NLP was an obvious fit for me when I started college, and insisting that I should still apply to grad school when I had my doubts at the end of college. I hope you would have been proud I finished this PhD.

Thank you as well to Andreas Stolcke and Jerry Feldman for the undergraduate research mentoring at ICSI. Thanks to Daniel Klein for teaching the rarely offered NLP class my senior year, and Greg Durrett for being a patient and enthusiastic TA.

Thank you to Phoebe Mulcaire for the company as we co-worked while applying to grad schools.

Thank you to my dear friends from Berkeley: Kate Rakelly, Amy Tsai, Madeeha Ghori, Armita Manafzadeh, and Jaya Narasimhan for the continued friendship from afar, and willingness to lend an ear. Kate was next to me from the very first day of KickStart in 2011, and was a diligent and disciplined study buddy and lab partner who also knew when it was time to take a Frisbee break. Thanks for continuing to be a dear friend from afar. Amy was by my side as we organized KickStart, and as we

learned our first bits of data science together. Thank you for the letters, they were a beautiful joy to receive in grad school. Thank you to Madeeha, for the listening, and the laughter.

Family

Last but most certainly not least, thank you to my parents. When I was a kid, my dad always said to, "stay in school as long as possible." For once, I listened. My mom has always been enthusiastic about my work. Thank you both for all your continued love and support.

Dedication

For my parents—

thank you, for everything.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

Imagine a world where everyone can access the information they need, no matter what language they speak.

Human translators play an important role in global communication, but they can only translate around 2,000 words a day (Chan, 2002). In contrast, current machine translation systems can translate that many words (or more) per second (Heafield et al., 2020).[1] Additionally, there are thousands of languages spoken worldwide; it is not always easy (or affordable) to find a human translator who can translate between particular languages, or who can translate a particular style of text. Machine translation (MT) is a subarea of natural language processing (NLP) that has the potential to fill the gaps to enable seamless communication across language barriers, and improve global information access.

However, for this to become a reality, MT must be available for all languages and styles of text.

Automatic translation has been a dream for decades, beginning with human-written translation rules applied by a computer. More modern approaches have treated machine translation as a machine learning problem, using existing human translations to *learn* high-output-dimensional structured-prediction translation models. Recent improvements in machine translation have made it more widely usable, partly due

[1] There are also computer assisted/aided translation (CAT) methods which use machine translation to assist human translators (Knowles and Koehn, 2016; Wuebker et al., 2016).

to deep neural network approaches. Neural machine translation is now deployed commercially, including in consumer facing applications by Microsoft, Google, and Facebook, among others. When trained on large high quality corpora, such models have even been shown to be near human quality in specific languages and domains where training data is abundant (Hassan et al., 2018).

However, machine translation is not yet effective for all settings and use-cases since—like most deep learning algorithms—neural machine translation is sensitive to the quantity and quality of the training data, and many of the situations where this technology is most needed lack large, high quality corpora. This is the focus of my work:[2] overcoming data challenges by improving machine translation in settings which lack sufficient high quality corpora.

1.2 Overview

Machine translation models translate a sentence in the source language to a sentence in the target language (e.g., translating from Spanish to English). Such models are typically trained on pairs of sentences that were originally translated by humans. The current state-of-the-art solution to this machine learning problem is neural machine translation (NMT), where models are deep neural networks with parameters estimated by training on the parallel training data.

[2]With the exception of the introduction and conclusion, the main body of this dissertation uses the first person plural ('we') rather than the singular ('I'). This is to both reflect standard practice in the field, and also to respect contributions made by collaborators in the case of joint work.

Limited quantities of such data are available for most language pairs, leading to a *low resource* problem. We present a method for transfer learning from a paraphraser to overcome data sparsity in low resource settings in Chapter 3.

Even when training data is available in the desired language pair, it is frequently formal speech or news—leading to a *domain* mismatch when models are used to translate a different type of data from most of what they were trained on. We present a method for improving domain adaptation translation quality in Chapter 4.

Neural machine translation currently performs poorly in domain adaptation and low resource settings (Koehn and Knowles, 2017; Sennrich and Zhang, 2019). A seemingly obvious approach when faced with a lack of data is to go get more data. This is often the best way to improve translation quality. However, it is not always feasible to produce additional human translations. In such a case, an option may be to crawl the web for additional training data. However, such data can be very noisy and harm machine translation quality— particularly neural machine translation quality—as we show in Chapter 5. We will also discuss some of the noise mitigation methods that were inspired by our work in Chapter 5.

1.3 Publications

Portions of this dissertation have been previously published, and additional work completed during my doctoral studies has also been published. Here I categorize and

briefly summarize the papers.

1.3.1 Low Resource Machine Translation

The overarching challenge in low resource machine translation is data sparsity. I have addressed this using a paraphraser by:

- Generating additional training data by paraphrasing one side of a parallel corpus (Hu, Khayrallah, Culkin, Xia, Chen, Post, and Van Durme, 2019a).
- Simulating training on many possible translations using a paraphraser in the training objective (Chapter 3; Khayrallah, Thompson, Post, and Koehn, 2020a).[3]

A specific challenge in low resource MT is vocabulary coverage; words in the text we would like to translate are often unobserved in the parallel training corpus. I developed methods to improve translation of rare and unknown words including:

- Morphological segmentation to improve statistical machine translation of rare words (Ding, Duh, Khayrallah, Koehn, and Post, 2016).
- Generation of lexical translations and integration of those translations in statistical machine translation (Gujral, Khayrallah, and Koehn, 2016).
- Morphological re-inflection to generate additional forms of words (Cotterell, Vylomova, Khayrallah, Kirov, and Yarowsky, 2017).

[3]This paper was nominated for best paper at WeCNLP 2020.

- Integration of lexical translations in statistical machine translation (Shearing, Kirov, Khayrallah, and Yarowsky, 2018).

- Integration of lexical translations in neural machine translation (Thompson, Knowles, Zhang, Khayrallah, Duh, and Koehn, 2019a).

1.3.2 Domain Mismatch

I have addressed the problem of domain mismatch by:

- Proposing a method for combining neural and statistical MT to reduce inaccurate translations that would confuse users, and applying it to domain adaptation (Khayrallah, Kumar, Duh, Post, and Koehn, 2017).

  - Applying Khayrallah et al. (2017) to low resource machine translation (Ding, Khayrallah, Koehn, Post, Kumar, and Duh, 2017).

- Adding a regularization term during adaptation that keeps the model output from differing too much from the original generic model, and improves performance in the domain of interest (Chapter 4; Khayrallah, Thompson, Duh, and Koehn, 2018a).

- Analyzing models during adaptation (Thompson, Khayrallah, Anastasopoulos, McCarthy, Duh, Marvin, McNamee, Gwinnup, Anderson, and Koehn, 2018).

- Reducing forgetting of original domain knowledge during adaptation (Thompson, Gwinnup, Khayrallah, Duh, and Koehn, 2019b).

1.3.3 Noisy Training Data

Towards addressing the challenge of noisy training data I have:

- Demonstrated that web-crawled training data can contain noise that degrades translation quality (Chapter 5; Khayrallah and Koehn, 2018).[4]

- Organized a shared task on filtering web-crawled data to address that noise problem (Koehn, Khayrallah, Heafield, and Forcada, 2018).

- Applied a method for filtering noisy data (Khayrallah, Xu, and Koehn, 2018b).

1.3.4 Additional NLP Contributions

In addition to the three focus areas of this dissertation, I have contributed to other areas of NLP, including:

- Creating an interface for teaching about machine translation (Khayrallah, Knowles, Duh, and Post, 2019).

- Developing a multiview learning method to incorporate multiple views of data.

  This work was originally motivated by improving word-embeddings for use in

[4]This paper won the Outstanding Contribution Award at the 2018 Workshop on Neural Machine Translation and Generation (WNMT).

bilingual lexicon induction to be used to translate out of vocabulary words, but it ended up being effective for phonetic transcription from acoustic & articulatory measurements, recommending hashtags, and recommending friends on a dataset of Twitter users (Benton, Khayrallah, Gujral, Reisinger, Zhang, and Arora, 2019).

- Generating a comprehensive list of translations (Khayrallah, Bremerman, McCarthy, Murray, Wu, and Post, 2020b).

- Applying simulated multiple reference training (Chapter 3; Khayrallah et al., 2020a) to non-task oriented dialog (Khayrallah and Sedoc, 2020).

- Proposing a linguistically motivated diagnostic for the 'I don't know' problem in non-task oriented dialog (Khayrallah and Sedoc, 2021).

Chapter 2

Background

In 2015,[1] statistical machine translation (SMT) was working well (Bojar et al., 2015), but a new paradigm (neural machine translation; Kalchbrenner and Blunsom, 2013) was becoming competitive (Jean et al., 2015a).

In some ways neural machine translation (NMT) revolutionized the field—it led to higher translation quality in high resource settings, and new perspectives on transfer learning (as applied in Chapters 3 and 4). This new paradigm also came with new challenges—different computer hardware requirements, less fidelity, and less robustness (as explored in Chapter 5). Additionally, some familiar challenges remained: low resource and domain mismatch settings as explored in Chapters 3 and 4.

As a background to the remaining chapters, we provide brief history of machine translation, followed by a high level introduction to the machine translation paradigms we will use in this thesis (SMT and NMT). Additionally, we will discuss how such systems are evaluated. Perhaps most importantly—particularly in the context of this thesis—we will discuss the role of different types of data in machine translation.

2.1 Historical Context

Automatic translation has been a dream for decades, if not longer. We review this history to not only provide perspective, but also to contextualize the current state of the art models, and highlight the fact that many of the current challenges are long standing ones.

[1] When I began my doctoral studies.

`babelfish.altavista.com`, was launched on December 9, 1997, named after a fictional idea of a fish that could be placed to one's ear and translate between languages (Adams and Perkins, 1985). This may have been the first publicly facing machine translation service.

But well before that, in a 1949 letter to Norbert Wiener, Warren Weaver commented (Hutchins, 1997):

> Also knowing nothing official about, but having guessed and inferred considerable about, powerful new mechanized methods in cryptography—methods which I believe succeed even when one does not know what language has been coded—one naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography. When I look at an article in Russian, I say "This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode."

This framing of the translation problem as code breaking has left its mark on the vocabulary of the field. 'Decoding' is often used to describe the process of generating a translation, and 'encoder' and 'decoder' are used to describe parts of neural machine translation models.

After a lack of optimism from Wiener, Weaver responded:

> Suppose we take a vocabulary of 2,000 words, and admit for good measure all the two-word combinations as if they were single words. The vocabulary is still only four million: and that is not so formidable a number to a modern computer, is it?

As we will discuss in Section 2.4.4, balancing the size of the vocabulary remains a challenge, even to 'modern' computers in 2021, over 70 years after the original correspondence.

2.2 Data Driven Machine Translation

Data driven machine translation consists of models, evaluation of those models, and data to train the models. We will describe each of these in subsequent sections.

Machine Translation is a structured prediction problem, with a high dimensional output space (vocabulary) and with many acceptable sequences (translations) for each example. There have been many different approaches to solving this problem. We will review the two data driven paradigms that will be referenced in later chapters of this thesis: statistical machine translation and neural machine translation.

We can describe the process of translation as finding the most likely target sentence y given a source sentence x.

This can be written mathematically as:

$$\arg\max_{y} p(y|x) \tag{2.1}$$

Statistical machine translation, and neural machine translation each have different ways of learning probability distribution from data.

2.3 Statistical Machine Translation

We give a high level overview of statistical machine translation, with a focus on phrase based machine translation (Koehn et al., 2003).[2] We provide this as background for Chapter 5; for a more thorough explanation, consider the textbook by Koehn (2009), or the book chapter by Osborne (2010).

Statistical machine translation learns to translate between a source and a target language from both parallel and monolingual corpora.

Using Bayes rule, we can break Equation 2.1 down into:

$$\arg\max_y p(y|x) = \arg\max_y \frac{p(x|y)p(y)}{p(s)} \tag{2.2}$$

However, when translating a given sentence x, $p(x)$ will always be the same, leaving:

$$\arg\max_y p(y|x) = \arg\max_y p(x|y)p(x) \tag{2.3}$$

Equation 2.3 is referred to as a noisy-channel model, and comes from information theory (Shannon, 1948). This brings us back to the analogy by Weaver; the noisy channel is modeling translation under the assumption that the sentence was intended to be in the target language, but got distorted in a noisy channel and ended up in the source language.

[2]We note that there are other statistical machine translation approaches such as hierarchical phrase-based models (Chiang, 2007) and syntax-based models (Galley et al., 2004; Galley and Manning, 2008)

The noisy-channel model requires $p(x|y)$ and $p(y)$ (and a search for the arg max). $p(y)$ is a language model, it predicts the probability of each word, given some previous words ($p(y_i|y_{j<i})$). This can be learned from monolingual text. N-gram language models are most typically used, and these can be learned from either the target side of the parallel text, or from additional monolingual data.[3]

$p(x|y)$ is the translation model. We will provide a high level overview, and note that modeling $p(x|y)$ is where the variety of statistical machine translation paradigms differ.

The translation model is trained on a parallel corpus.[4] Based on that parallel corpus, an alignment is learned between words in the source and target sentences. Then phrasal translations are extracted based on those alignments. Note that since alignment is done within a sentence pair, a phrase translation cannot be extracted if the source and target phrase do not occur in an aligned sentence pair.

In practice, a variety of features go in the model, weights on which are then learned during tuning.

'Decoding,' or generating a translation, is a search process. For each hypothesis, the target sentence is typically generated phrase by phrase, in order,[5] but not necessarily in order of the source sentence; phrases from the source sentence can be translated in any order. A phrase is selected from the input sentence, and then a translation

[3]Monolingual data is typically easier to acquire, and increasing the amount for language model training typically improves translation quality, especially since it might be more domain relevant.

[4]Finding aligned documents and extracting aligned sentences are also steps in the data gathering pipeline, but are beyond this summary.

[5]e.g., 'left-to-right' for English.

is chosen for it. In addition to the probability of that translation, language model probabilities and other feature scores are combined for scoring. Since the problem is NP-hard, beam search is typically used.

2.4 Neural Machine Translation

Neural machine translation (Kalchbrenner and Blunsom, 2013) describes a variety of approaches that use 'end to end' neural networks for translation. These models are typically trained on pairs of parallel sentences[6] and use backpropagation of a loss to learn the weights of the neural network (Rumelhart et al., 1986).

The first successful approaches to neural machine translation used encoders and decoders (Sutskever et al., 2014; Cho et al., 2014), typically based on recurrent neural networks or variants. The introduction of attention allowed for a focus on different parts of the full input sequence and improved translation quality (Bahdanau et al., 2015). The Transformer model, introduced by Vaswani et al. (2017), addresses the recency bias of recurrent neural networks by forgoing recurrent connections in favor of more attention throughout the model.

Regardless of the architecture of the models, neural machine translation models typically produce tokens one by one, and typically generate the target sentence in order.[7] This is done by taking the softmax over the size of the vocabulary, and

[6] Though document level approaches have been explored (e.g., Junczys-Dowmunt, 2019).
[7] e.g., 'left-to-right' for English

choosing the highest probability token.[8]

During standard neural machine translation training, the reference is: (1) used in the training objective; and (2) conditioned on as the target prefix.[9]

2.4.1 NLL Objective

Neural machine translation models are typically trained using negative log likelihood (NLL) with respect to a single reference. The standard negative NLL training objective in NMT, for the i^{th} target word in the reference y is:

$$\mathcal{L}_{\text{NLL}} = -\sum_{v \in \mathcal{V}} \Big[\mathbb{1}\{y_i = v\} \times \log\ p_{\text{MT}}(y_i = v \,|\, x, y_{j<i}) \Big] \tag{2.4}$$

where $\mathcal{V}$ is the vocabulary, $\mathbb{1}\{\cdot\}$ is the indicator function, and p_{MT} is the MT output distribution (conditioned on the source x, and on the previous tokens in the reference $y_{j<i}$). Equation 2.4 computes the cross-entropy between the MT model's distribution and the one-hot reference.

[8] Beam search can also be used.

[9] In autoregressive NMT inference, predictions condition on the previous target tokens. In training, predictions typically condition on the previous tokens in the reference, not the model's output (teacher forcing; Williams and Zipser, 1989).

2.4.2 Knowledge Distillation

An alternative approach is (word-level) Knowledge Distillation (Hinton et al., 2015; Kim and Rush, 2016) which assumes access to a teacher distribution ($p_{teacher}(y \mid x)$) and minimizes the cross entropy with the teacher's probability distribution.

The knowledge distillation training objective for the i^{th} target word in the reference y, given the source x, with a target vocabulary $\mathcal{V}$ is:

$$\mathcal{L}_{\text{KD}} = -\sum_{v \in \mathcal{V}} \Big[p_{\text{TEACHER}}(y_i = v \mid y, y_{j<i}) \times \log \left(p_{\text{MT}}(y_i = v \mid x, y_{j<i}) \right) \Big]$$

The teacher and student each condition on the previous reference tokens ($y_{j<i}$).

Kim and Rush (2016) introduced sequence-level knowledge distillation. This re-frames knowledge distillation as a data augmentation problem: the teacher model is used to generate full sequences, which are then paired with the original source to form a parallel corpus which the child is trained on (using NLL).

2.4.3 Regularization

Regularization is important part of machine learning models to prevent overfitting. Examples used in neural machine translation include: dropout and label smoothing.

Srivastava et al. (2014) propose dropout to prevent overfitting in neural networks.

This technique randomly selects some nodes to be ignored or 'dropped out' during training, forcing other nodes to adapt.

Label smoothing spreads some probability mass over all non-reference tokens equally (Szegedy et al., 2016). It can be viewed as a weighted average between the (one-hot) gold target, and the uniform distribution over all the labels (typically with a larger weight on the gold target). Müller et al. (2019) analyze label smoothing and find that it not only improves generalization, but also improves model calibration, which in turn improves beam search.

2.4.4 Subword Vocabularies

One of the bottlenecks of machine translation is taking the softmax over the target vocabulary; this is slow. Additionally, word embeddings are a large percentage of the memory used by the model.

For these reasons, early neural machine translation models limited their vocabularies to a fixed size,[10] replaced rare words with an 'unk' token, and then backed off to a dictionary (Jean et al., 2015b; Luong et al., 2015).

An alternative approach is to break up rarer words into smaller units. Morphological segmentation and compound splitting were explored for statistical machine translation (e.g., Nießen and Ney, 2000; Koehn and Knight, 2003; Virpioja et al., 2007; Stallard et al., 2012).

[10] This size tended to range from 30K to 100k tokens.

The first widely used word segmentation approach for neural machine translation—proposed by Sennrich et al. (2016c)—adapted the byte pair encoding (BPE; Gage, 1994) compression algorithm for this task. Byte pair encoding iteratively merges the most frequently adjacent pair of bytes and replaces them with a new byte. Sennrich et al.'s adaptation for NMT begins by initializing the vocabulary with all characters found in the text. The most frequently occurring pair of characters (perhaps 't', 'h' in English) is replaced with a new symbol (e.g., 'th') and this new symbol is added to the vocabulary. This process continues until a preset maximum number of merges is conducted. The final vocabulary is the initial character set, plus the new symbols created by merges. In practice, merges are not allowed across word boundaries for efficiency. When the segmentation is applied to text, it is indicated by a special marking allowing the segmentation to be removed.[11] Chapters 4 and 5 use BPE.

In 2018 another segmentation algorithm (Kudo, 2018) was proposed, and released as part of the SentencePiece tool kit (Kudo and Richardson, 2018). While BPE assumes that the data has been tokenized into words,[12] SentencePiece does not have that assumption, and also does not require tokenization as a preprocessing step.[13]

[11] e.g. 'underneath' might get segmented as under@@ ne@@ ath. The original can be recovered with the sed command `sed -r 's/(@@ )|(@@ ?$)//g'`.

[12] While this assumption is somewhat reasonable for some languages such as English which separate words with white space, spaces are not required in some other languages, such as Japanese and Chinese.

[13] SentencePiece treats space as a 'character' by replacing it with a special symbol ('▁'). Segmentation can be removed with the python command `detokenized = ''.join(segmented).replace('▁', ' ')`.

SentencePiece releases a variety of approaches. In Chapter 3, we use the unigram model, which assumes that the probability of a sequence of subwords is the product of the probability of those subwords. Finding the most probable segmentation is then an argmax, which can be computed using the Viterbi algorithm (1967). Straightforward EM is not possible in this case, so a modified iterative algorithm is used to learn the vocabulary and probabilities.

Chitnis and DeNero (2015) proposed a word segmentation model for neural machine translation using Huffman encoding (1952), but unlike the approaches of Sennrich et al. (2016c) and Kudo (2018), the segmentation based on Huffman encoding does not produce symbols that are interpretable as subword units, and cannot generalize to translate and produce new words unobserved at training time.

While early NMT models used subword vocabularies as large as possible (e.g., 30-100k tokens), there is evidence that smaller vocabulary sizes improves translation quality, particularly for lower resource settings (Ding et al., 2019), a trend followed by Guzmán et al. (2019) and we follow in Chapter 3. Using a smaller vocabulary means even relatively common words will be segmented during training. This reduces sparsity and may allow the model to learn how to translate rare variants of those words. It is important to note that these automatic segmentation algorithms do not aim to explicitly learn morphologically plausible subword units.[14]

[14]In Ding et al. (2016), we explore both BPE segmentation and morphological segmentation for SMT.

2.5 Comparing Statistical and Neural Machine Translation

Statistical machine translation translates discrete tokens explicitly, and a word can only be generated by the model if it is part of a phrase pair that also occurred in a parallel sentence pair in the training data. This level of fidelity does not apply in neural machine translation. While neural machine translation models do perform better in general—in part due to their ability to generalize—this allows them to 'hallucinate' (generate output unrelated to the input). Unlike inadequate translations in statistical machine translation—which often take the form of disfluent outputs—neural machine translation errors are often fluent in the target language, making them difficult to identify by a monolingual speaker (Martindale and Carpuat, 2018; Martindale et al., 2019; Martindale, 2020). As we will discuss in Chapter 5, NMT struggles with robustness to certain types of noise—both in training and in decoding.

A variety of approaches combined neural and statistical machine translation in hybrid systems to balance the benefits of each paradigm (e.g., Devlin et al., 2014; Junczys-Dowmunt et al., 2016; Stahlberg et al., 2016; Mi et al., 2016; Stahlberg et al., 2016; Stahlberg et al., 2017; Khayrallah et al., 2017).

Despite producing higher quality translations than statistical machine translation in high resource single domain settings, initial neural machine translation models under-performed statistical machine translation models in several difficult data

conditions such as translating under domain mismatch, in low resource settings, when translating rare words (Koehn and Knowles, 2017), and noisy data settings (Khayrallah and Koehn, 2018). However, more recent work has mitigated this reduction in translation quality. Neural machine translation does have properties which can make it advantageous in some of these settings: e.g, the segmentation of words into subword units (which may allow for learning of different morphological variants) and the transfer learning approaches neural machine translation enables.

2.6 Evaluation

While the gold-standard for evaluation of machine translation models is human evaluation, that is not always feasible.[15] There are a variety of different automatic metrics for machine translation that compare the similarity between the machine translation model's output and a human reference translation.[16] This similarity can be judged in a variety of ways.[17] The current standard is the BLEU score (Papineni et al., 2002), which is a weighted n-grams precision between the machine translation output and human reference:

[15] In addition to potentially being expensive, human evaluation cannot be directly optimized towards, and is not consistent (two different evaluators may give different responses).

[16] While we still depend on a human reference, this can be reused to evaluate many different systems, rather than just one.

[17] e.g.: Papineni et al. (2002), Doddington (2002), Lavie and Agarwal (2007), Lo and Wu (2011), Stanojević and Sima'an (2014), Gupta et al. (2015), Gupta et al. (2015), Popović (2015), Popović (2017), Lo (2017), Shimanaka et al. (2018), Tiedemann and Scherrer (2019), Mathur et al. (2019), Lo (2019), Chow et al. (2019), Yankovskaya et al. (2019), Zhang et al. (2020), Sellam et al. (2020), and Thompson and Post (2020a).

$$BLEU_4 = min(1, \frac{\text{output length}}{\text{reference length}}) \prod_{i=1}^{4} precision_i \tag{2.5}$$

Where $precision_i$ is the precision of i-grams (e.g. the ratio of correct i-word phrases to the total number of i-grams in the machine translation output.) This gives a score between 0 and 1, which is typically scaled to be between 0 and 100 for readability. BLEU is typically computed over a corpus, rather than on a sentence level.

While BLEU is beginning to show its age and may not be ideal for comparing extremely similar quality systems—e.g., very high quality systems (Ma et al., 2019; Mathur et al., 2020)—it currently remains the standard metric after nearly two decades. It is important to note that BLEU scores can only be directly compared on a single test set, and cannot be compared across languages. Since BLEU computes n-gram matches, tokenization influences the score, and BLEU can also only be directly compared with consistent tokenziation. SacreBLEU (Post, 2018)[18] is a package that re-implements the tokenization of `mteval-v13a.pl`, the official script of the Conference on Machine Translation (WMT) evaluations.[19]

[18] github.com/mjpost/sacrebleu
[19] github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/mteval-v13a.pl

2.7 Data

Like most machine learning algorithms, data-driven machine translation (including both SMT and NMT) typically requires a 'training' data set, and a 'test' set for reporting results. The former should be as large as possible, and the latter is typically on the order of 1,000 to 3,000 sentences. Statistical machine translation uses a tuning set to learn the weights of different features. Neural machine translation uses a development (dev) set for model selection from checkpoints or for early stopping. Both tuning and development sets are typically the size of test sets.

Machine translation models are typically trained on a parallel corpus, which consists of pairs of sentences originally translated by human translators[20] although other forms of data can be incorporated in training as well. Table 2.1 shows a toy example of a parallel corpus.[21]

La liebre y la tortuga.	The hare and turtle.
La tortuga verde.	The turtle is green.
El conejo tiene orejas.	The rabbit has ears.
Una liebre rápida.	A fast hare.

Table 2.1: Example Spanish-English parallel training data.

Recent improvements in machine translation modeling have made it more widely usable, however, translation quality still heavily depends on data quality and quantity.

An obvious solution to problems of data scarcity is to get more data. This can

[20]The process of extracting these sentences is known as sentence alignment. See Koehn (2009) for a description of the problem, and Thompson and Koehn (2019) for a recent approach.

[21]This can also be referred to as bitext, parallel text, or parallel data.

come in the form of transfer learning, data augmentation, and gathering additional parallel data from the web.

2.7.1 Transfer Learning

We can generalize the problem of domain adaptation to one where there is a small relevant parallel corpus, and a large less relevant corpus.

Domain adaptation can be seen as a kind of low resource setting, where there is insufficient data in the language pair *and domain* of interest (although there may be plenty of data in the language pair in general). Some similar approaches can therefore be applied to low resource and domain adaption settings. Transfer learning across different domains as well as languages and/or dialects is now common in NLP for low resource settings.

2.7.1.1 Continued Training

A simple yet effective technique commonly applied in adaptation settings is continued training[22] (Luong and Manning, 2015), where a model is first trained on the larger general corpus, and then that model is used to initialize a new model that is trained on the more specific corpus.

Continued training was initially proposed for domain adaptation (Luong and

[22]This can also be referred to as *fine tuning*, we use the term *continued training* to distinguish from the framework of Hinton and Salakhutdinov (2006), which uses supervised learning to fine tune features obtained through unsupervised learning (and for consistency with the notation in the published version of Chapter 4).

Manning, 2015), but can also be applied to other forms of transfer learning. For consistency with the literature, we will describe continued training using the terminology of domain adaptation. We will then discuss how this can be used as transfer learning in other types of low resource settings.

Continued training consists of three steps:

1. Train a model until convergence on a large out-of-domain parallel corpus using $\mathcal{L}_{\text{NLL}}$ as the training objective.

2. Initialize a new model with the final parameters of Step 1.

3. Train the model from Step 2 until convergence on in-domain parallel corpus, again using $\mathcal{L}_{\text{NLL}}$ as objective.

In other words, continued training initializes an in-domain model training process with parameters from an out-of-domain model. The motivation is that the out-of-domain model provides a reasonable starting point and is better than random initialization.[23]

2.7.1.2 Continued Training for Domain Adaptation

Empirically, continued training works very well for domain adaptation, and there are several variants. For example, in Chapter 4 we introduce a regularization technique for continued training of machine translation models that improves translation

[23] In Thompson et al. (2018) we show evidence for this hypothesis.

quality in domain adaptation. This keeps the model output from differing too much from the original general model, and improves translation quality in the domain of interest. During standard fine-tuning, in-domain improvements from adaptation come at the expense of general-domain translation quality; this method mitigates the domain-adapted model's drop in translation quality on the original domain. Dakwale and Monz (2017) use a similar approach but focus on preventing the domain-adapted model's drop in translation quality on the original domain rather than improving adaptation translation quality. In Thompson et al. (2019b), we interpret this drop in general-domain translation quality during standard fine-tuning as catastrophic forgetting. To mitigate it, we adapt elastic weight consolidation (a machine learning method for combating catastrophic forgetting) to retain the majority of general-domain translation quality lost without degrading in-domain translation quality.

2.7.1.3 Additional Approaches to Domain Adaptation

There are additional non-continued training domain adaptation techniques. Some of them could be combined with continued training.

Instance weighting was originally proposed for domain adaptation in statistical NLP (Jiang and Zhai, 2007) and applied widely for statistical machine translation (e.g., Matsoukas et al., 2009; Shah et al., 2010; Foster et al., 2010; Rousseau et al., 2011; Zhou et al., 2015; Wang et al., 2016; Imamura and Sumita, 2016). This method scores each sentence or domain, and then trains the model with that score as the

weight on the sentence or domain. Wang et al. (2017) apply instance weighting to neural machine translation, and also propose a dynamic weight learning strategy.

Kobus et al. (2017) propose domain control for NMT to create a single model that can perform well on multiple domains. That work aims to provide the NMT encoder with meta-information about the domain, to allow it to learn to translate multiple different domains well. They propose two methods: (1) a domain specific token added to source sentence (inspired by Sennrich et al., 2016a), and (2) a domain embedding portion added to the word embeddings (inspired by Crego et al., 2016).

In Khayrallah et al. (2017), we use the lattice output of statistical machine translation to constrain the search space available to a neural machine translation decoder, bringing together the robust adequacy and the fluency properties of statistical MT and neural MT systems, respectively. Incorrect translations which read fluently in the target language but are unrelated to the original source sentence were a problem in early neural machine translation systems, especially in domain mismatch settings. Such translations are particularly problematic because the person reading them might not realize they are incorrect since they read so fluently.

For a survey of domain adaptation that includes both continued-training and additional approaches, see (Saunders, 2021).

2.7.1.4 Analysis of Domain Adaptation

In addition to work that aims to directly improve domain adaptation, there is work that aims to analyze the domain mismatch problem. In Thompson et al. (2018), we analyze different components of the neural network to better understand what happens during adaptation. We find that the models are still able to adapt well when any single part of the model remains fixed, and that while training on general domain data alone does not lead to good translation quality, it does get the model close to a good local minimum in the in-domain error surface, making it well placed for adaptation on the in-domain data.

Gu and Feng (2020) perform a similar analysis, though they use the transformer architecture (Vaswani et al., 2017), and focus their analysis on the problem of catastrophic forgetting in NMT[24] and find different parts of NMT models are important for general and in-domain translation quality.

2.7.1.5 Crosslingual Transfer

Zoph et al. (2016) apply continued training to transfer between high and low resource language pairs to improve low resource translation quality. They experiment with language pairs of different levels of similarity, e.g., transferring to Uzbek–English from French-English, and transferring to Spanish-English from both French–English and German–English. They find that transfer from French–English to Spanish-English

[24] A problem we touch on in Chapter 4, and we address in Thompson et al. (2019b).

performs better than transferring from German–English to Spanish-English. They also experiment with freezing various parts of the model when transferring from French–English to Uzbek–English. They find freezing target embeddings and training all other parameters works best.

Zoph et al. (2016) do not use any subword units,[25] and they initialize input language embeddings for the child model with randomly-assigned embeddings from the parent.[26] Nguyen and Chiang (2017) learn a BPE subword vocabulary on the combined source and target data of both the parent and child languages. They consider the case of transfer between related languages (e.g., in the Turkic family). They find that even though some of the languages may be written in different scripts (in which case they apply transliteration as a preprocessing step), after applying BPE there is an over 50% vocabulary overlap in the training data. While the word-based transfer method of Zoph et al. does not always improve translation quality in Nguyen and Chiang's experiments, the BPE-based transfer does.

Dong et al. (2015) and Firat et al. (2016) consider multilingual neural machine translation models, and the transfer that can occur between languages pairs. These models use a different encoder or decoder for each language.

Firat et al. (2016) find the transfer between languages particularly helpful in the simulated lower resource pairs they considered, though the smallest setting they consider is 100k lines of training data.

[25] Zoph et al. (2016) was published less than three months after Sennrich et al. (2016c).

[26] All the experiments use English as the target language (both parent and child models), so no change is required to the target side embeddings.

Johnson et al. (2017) consider a simpler architecture, using tags to indicate the language (inspired by Sennrich et al., 2016a). They find it improves low resource translation. They also explore zero-shot translation—translation between two languages that were trained on as part of other pars, but had no parallel corpus between them (e.g., a model trained on Portuguese-English and English-Spanish translation can generate reasonable translations for Portuguese-Spanish). Additionally, they find that zero-shot translation can be improved by continued training on a small amount of language-pair specific data.

2.7.1.6 Pretraining on Monolingual Data

Large pretrained encoder models trained on monolingual data (as opposed to a parallel corpus), sparked by the success of ELMo (Peters et al., 2018), have revolutionized NLP (for a survey, see Xia et al., 2020). Some of these have variants that are trained on monolingual data in multiple languages. Many of these are simply encoders, and are often used to generate contextual embeddings.

BART (Lewis et al., 2020) is a sequence to sequence transformer, and mBART was proposed as a multilingual version of BART as a method for pretraining MT (Liu et al., 2020). It is trained on parallel corpora synthetically generated from noised monolingual data in multiple languages. The target sentence is the original sentence, and the input is a noised version of that sentence. The types of noise are:

- Token Masking: random tokens are sampled and replaced with a [MASK] token

(Devlin et al., 2019).

- Token Deletion: random tokens are deleted.
- Text Infilling: masking multiple tokens with a single mask (inspired by Joshi et al., 2020).
- Sentence Permutation: randomly shuffling phrases in the sentence.
- Document Rotation: choosing a random token as the start, and keeping the rest in order (wrapping around the text).

Liu et al. found that this pretraining improves translation quality at all but the highest resource levels (over 25 million lines).

2.7.2 Data Augmentation

Ideally, we would like to have larger quantities of (high-quality) data to train. However, there is typically a limit to the amount of human-translated data available for any given language pair and domain. Data augmentation is a family of approaches that create additional synthetic training data, often (though not always) based on monolingual data.

Back-translation

Back-translation (Sennrich et al., 2016b) is the most common method for data augmentation using non-parallel data in NMT. Back-translation translates

target-language monolingual text to create synthetic source sentences. Back-translation requires a reverse translation model for each language pair but is effective at a variety of resource levels. Additionally, it can be effective at incorporating monolingual domain-specific text for adaptation.

There are several of variants of back-translation. Fadaee and Monz (2018) select sentences to back translate which have (1) difficult words, or (2) difficult contexts for such words. Edunov et al. (2018) propose sampled back-translation and found that sampling when generating the back-translations improved translation quality. Caswell et al. (2019) propose tagged back-translation, which signals to the model which sentence pairs are synthetic using tags (inspired by Sennrich et al., 2016a). Iterative back-translation iteratively trains machine translation models in the source-target and target-source directions, and improves each of them with back-translation repeatably (Hoang et al., 2018).

Additional Approaches

Fadaee et al. (2017) insert rare words in novel contexts in the existing parallel corpus, using automatic word alignment and a language model. RAML (Norouzi et al., 2016) and SwitchOut (Wang et al., 2018b) randomly replace words with others from the vocabulary during training.

Currey et al. (2017) train multitask machine translation models that learn to both translate source language text and copy target language text. They do so by creating

synthetic parallel corpora by copying monolingual target language data to the source, and mixing that with the parallel training data. This improves translation quality for words that should be identical in both languages (e.g., named entities).

2.7.3 Web-crawled Data

Even with improved methods and models, the tried-and-true method for improving translation is gathering more data.

One approach to complement transfer learning, data augmentation, and the often prohibitively expensive task of having translators translate millions of sentences for model training is to crawl the web for existing translated data.

Although the idea of crawling the web for parallel data goes back to the 20th century (Resnik, 1999), work in the academic community on extraction of parallel corpora from the web mostly focused on large stashes of multilingual content in (relatively) straightforward to align form, such as the Canadian Hansards, Europarl (Koehn, 2005), the United Nations (Rafalovitch and Dale, 2009; Ziemski et al., 2016), or European Patents (Täger, 2011). A curated product of these efforts is the OPUS web site (Tiedemann, 2012; Skadiņš et al., 2014).[27]

Paracrawl is an ongoing large-scale effort to crawl text from the web (Bañón et al., 2020). Acquiring parallel corpora from the web typically goes through stages of: (1) identifying web sites with parallel text, (2) downloading the pages of the web

[27] opus.nlpl.eu

site, (3) aligning document pairs, and (4) aligning sentence pairs. A final stage of the processing pipeline (5) filters out bad sentence pairs. These bad sentence pairs exist either because the original web site did not have any actual parallel data, only partial parallel data, or due to failures of earlier processing steps.

As we show in Chapter 5, unfiltered crawled data degrades translation quality. To encourage more research on this challenge, we organized a shared task on filtering web-crawled data (Koehn et al., 2018).

Chapter 3

Improving Low-Resource Machine Translation with Simulated Multiple Reference Training

3.1 Introduction

As discussed in Chapter 1, there are a variety of low resource settings that lack sufficient training data to build high quality machine translation models. In this chapter,[1] we introduce a method for transfer learning from a paraphraser in order to simulate having more parallel training data, in the form of multiple references per training example.

Many possible valid translations typically exist for a given sentence; in fact Dreyer and Marcu (2012) showed that naturally occurring sentences can have *billions* of valid translations. Despite this variety, machine translation models are optimized toward a single translation of each sentence in the training corpus. We hypothesize that the discrepancy between linguistic diversity and standard single-reference training hinders machine translation quality. Training a high resource MT model on millions of sentence pairs likely exposes it to similar sentences translated different ways, but training a low-resource MT model with a single translation for each sentence (out of potentially billions) exacerbates data sparsity.

This discrepancy was previously impractical to address, since obtaining multiple human translations of training data is typically not feasible. However, recent neural sentential paraphrasers produce fluent, meaning-preserving English paraphrases. We introduce Simulated Multiple Reference Training (SMRT), a method that incorporates

[1]The work described in this chapter was published in Khayrallah et al. (2020a). In Khayrallah and Sedoc (2020), we apply this method to non-task-oriented dialog systems (chatbots) and analyze the effect on response diversity.

such a paraphraser directly in the training objective, and uses it to simulate the full space of translations. SMRT approximates the full space of possible translations by *sampling* a paraphrase of the reference sentence from a paraphraser and training the MT model to predict the paraphraser's *distribution* over possible tokens.

We demonstrate the effectiveness of our method on two corpora from the low-resource MATERIAL program, and on parallel corpora from GlobalVoices.

We also analyze our method to understand:

1. how it performs at various resource levels;
2. how it combines with back-translation;
3. how the different components of the method impact translation quality; and
4. how it compares to sequence-level paraphrastic data augmentation.

3.2 Method

We propose Simulated Multiple Reference Training (SMRT), which uses a paraphraser to approximate the full space of possible translations, since explicitly training on billions of possible translations per sentence is intractable.

In standard neural MT training, the reference is:

1. used in the training objective; and

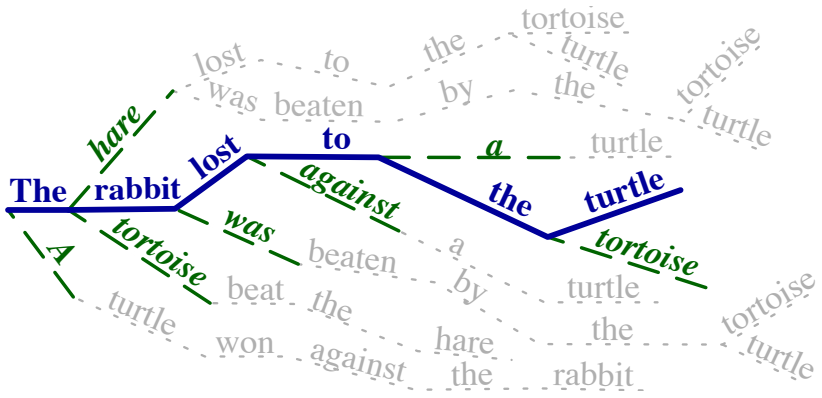
2. conditioned on as the target prefix.[2]

We approximate the full space of possible translations by:

1. training the MT model to predict the *distribution* over possible tokens from the paraphraser at each time step; and

2. *sampling* the previous target token from the paraphraser distribution.

Figure 3.1 shows an example of possible paraphrases and highlights a sampled path and some of the other tokens used in the training objective distribution.



Some possible paraphrases of the original reference, 'The tortoise beat the hare,' for the Dutch source sentence, 'De schildpad versloeg de haas.' A **sampled path** and some of the other ***tokens also considered in the training objective*** are highlighted.

Figure 3.1: A paraphrase example.

We review the standard NLL training objective, and then introduce our proposed objective.

[2]In autoregressive NMT inference, predictions condition on the previous target tokens. In training, predictions typically condition on the previous tokens in the reference, not the model's output (teacher forcing; Williams and Zipser, 1989).

3.2.1 NLL Objective

The standard negative log likelihood (NLL) training objective in NMT, for the i^{th} target word in the reference y is:

$$\mathcal{L}_{\text{NLL}} = -\sum_{v \in \mathcal{V}} \Big[\mathbb{1}\{y_i = v\} \times \log\ p_{\text{MT}}(y_i = v \,|\, x, y_{j<i}) \Big] \tag{3.1}$$

where $\mathcal{V}$ is the vocabulary, $\mathbb{1}\{\cdot\}$ is the indicator function, and p_{MT} is the MT output distribution (conditioned on the source x, and on the previous tokens in the reference $y_{j<i}$). Equation 3.1 computes the cross-entropy between the MT model's distribution and the one-hot reference.

3.2.2 Proposed Objective

In this work, rather than training towards that single one-hot reference, we would like to be able to train towards the full space of possible translations. We will do so by:

1. training the MT model to predict the *distribution* over possible tokens from the paraphraser at each time step (rather than the single one-hot vector y);
2. *sampling* a token from that distribution to use in the target prefix for both the MT model, and for the paraphraser.

We compute the cross entropy between the distribution of the MT model and the distribution from a paraphraser conditioned on the original reference:

$$\mathcal{L}_{\text{SMRT}} = -\sum_{v \in \mathcal{V}} \Big[p_{\text{para}}(y'_i = v \,|\, y, y'_{j<i}) \times \log\ p_{\text{MT}}(y'_i = v \,|\, x, y'_{j<i}) \Big] \tag{3.2}$$

where y' is a paraphrase of the original reference y. p_{para} is the output distribution from the paraphraser[3] (conditioned on the reference y and the previous tokens in the sentence produced by the paraphraser $y'_{j<i}$). p_{MT} is the MT output distribution (conditioned on the source sentence, x and the previous tokens in the sentence produced by the paraphraser, $y'_{j<i}$). At each time step we sample a target token y'_i from the paraphraser's output distribution to cover the space of translations. We condition on the sampled y'_{i-1} as the previous target token for both the MT model and paraphraser.

For a color-coded visualization see Figure 3.1, which shows possible paraphrases of the reference, 'The tortoise beat the hare.' The paraphraser and MT model condition on the **paraphrase (y')** as the previous output. The **paraphrase (y')** and the rest of the ***tokens in the paraphraser's distribution*** make up p_{PARA}, which is used to compute $\mathcal{L}_{\text{SMRT}}$.

[3]Paraphraser parameters are frozen during MT training.

3.3 Experimental Setup

3.3.1 Paraphraser

For use as an English paraphraser,[4] we train a Transformer model (Vaswani et al., 2017) in FAIRSEQ (Ott et al., 2019) with an 8-layer encoder and decoder, 1024 dimensional embeddings, 16 encoder and decoder attention heads, and 0.3 dropout. We optimize using Adam (Kingma and Ba, 2015). We train on PARABANK2 (Hu et al., 2019b), an English paraphrase dataset.[5] PARABANK2 was generated by training an MT system on CzEng 1.7 (a Czech−English parallel corpus with over 50 million lines (Bojar et al., 2016)), re-translating the Czech training sentences, and pairing the English output with the original English translation. Many potential candidates were generated from the translation model for each sentence, and high quality diverse paraphrases were selected.

3.3.2 NMT Models

We train Transformer NMT models in FAIRSEQ[6] using the FLORES low-resource benchmark parameters (Guzmán et al., 2019): 5-layer encoder and decoder, 512-dimensional embeddings, and 2 encoder and decoder attention heads. We regularize with 0.2 label smoothing and 0.4 dropout. We optimize using Adam with a learning

[4]We release paraphraser, the data and the code for replication: data.statmt.org/smrt
[5]Hu et al. released a trained SOCKEYE paraphraser but we implement our method in FAIRSEQ.
[6]We release paraphraser, the data and the code for replication: data.statmt.org/smrt

rate of 10^{-3}. We train for 200 epochs, and select the best checkpoint based on validation set perplexity. We translate with a beam size of 5. For our method we use the proposed objective $\mathcal{L}_{\text{SMRT}}$ with probability $p = 0.5$ and standard $\mathcal{L}_{\text{NLL}}$ on the original reference with probability $1 - p$. We sample from only the 100 highest probability vocabulary items at a given time step when sampling from the paraphraser distribution to avoid very unlikely tokens (Fan et al., 2018).

We use Tagalog (tl) to English (en) and Swahili (sw) to English parallel corpora from the MATERIAL low-resource program (Rubino, 2018). We also report results on MT parallel corpora from GlobalVoices, a non-profit news site that publishes in 53 languages.[7] We evaluate on the 10 lowest-resource settings that have at least 10,000 lines of parallel text with English: Hungarian (hu), Indonesian (id), Czech (cs), Serbian (sr), Catalan (ca), Swahili (sw),[8] Dutch (nl), Polish (pl), Macedonian (mk), and Arabic (ar).

We use 2,000 lines each for a validation set for model selection from checkpoints and for a test set for reporting results. The approximate number of lines of training data is in the top of Table 3.1. We train an English SentencePiece model (Kudo and Richardson, 2018) on the paraphraser data, and apply it to the target (English) side of the MT parallel corpus, so that the paraphraser and MT models have the same output vocabulary. We also train SentencePiece models on the source-side of the parallel corpus. We use a subword vocabulary size of 4,000 for each.

[7] We use v2017q3 released on Opus (`opus.nlpl.eu/GlobalVoices.php`; Tiedemann, 2012).

[8] Swahili is in both MATERIAL and GlobalVoices. MATERIAL data is not widely available, so we separate them to keep out GlobalVoices results reproducible.

3.4 Results

Results are shown in Table 3.1. Our method improves over the baseline in all settings, by between 1.2 and 7.0 BLEU (all statistically significant at the 95% confidence level (Koehn, 2004)).[9] We see larger improvements for lower-resource corpora.

dataset	GlobalVoices										MATERIAL	
* → en	hu	id	cs	sr	ca	sw	nl	pl	mk	ar	sw	tl
train lines	8k	8k	11k	14k	15k	24k	32k	40k	44k	47k	19k	46k
baseline	2.3	5.3	3.4	11.8	16.0	17.9	22.2	16.0	27.0	12.7	37.8	32.5
this work	**5.4**	**12.3**	**6.6**	**16.1**	**20.0**	**20.5**	**24.8**	**18.0**	**28.2**	**14.9**	**39.0**	**33.7**
Δ	+3.1	+7.0	+3.2	+4.3	+4.0	+2.6	+2.6	+2.0	+1.2	+2.2	+1.2	+1.2

Table 3.1: BLEU scores on the test set. We **bold** the best value; all improvements are statistically significant at the 95% confidence level. 'train lines' indicates the size of parallel corpus used for training.

3.5 Analysis

We analyze our method to explore:

1. how it performs at various resource levels (Section 3.5.1);
2. how it combines with back-translation (Section 3.5.2);
3. how the different components of the method impact translation quality (Section 3.5.3); and

[9] All BLEU scores are SacreBLEU (Post, 2018).

4. how it compares to sequence-level paraphrastic data augmentation (Section 3.5.4).

3.5.1 MT Data Ablation

In order to better understand how our method performs across various data sizes subselected from the same corpus, we ablate a Bengali-English parallel corpus from GlobalVoices.[10] After reserving data for evaluation, as in Section 3.3.2, approximately 132k lines are left for training; we ablate this to 100k, 50k, 25k, and 15k lines.

Figure 3.2 plots the translation quality of our method and the baseline against the log of the data amount. Our improvements of 2.7, 3.7, 1.6, and 0.8 BLEU at the 15k, 25k, 50k, and 100k subsets are statistically significant at the 95% confidence level; the 0.1 improvement for the full 132k data amount is not. Similar to Table 3.1, we see larger improvements in lower-resource ablations.

3.5.2 Back-translation

Back-translation (Sennrich et al., 2016b) is the most common method for incorporating non-parallel data in NMT. Similar to our work, it generates additional training data based on an auxiliary sequence-to-sequence model. It is a very effective form of data augmentation, so we investigate how our method interacts with it.

[10]We choose bn-en for its relatively large size while still containing dissimilar languages, as ablating French-English (another similarly-sized option from GlobalVoices) does not reflect typical low-resource machine translation quality.

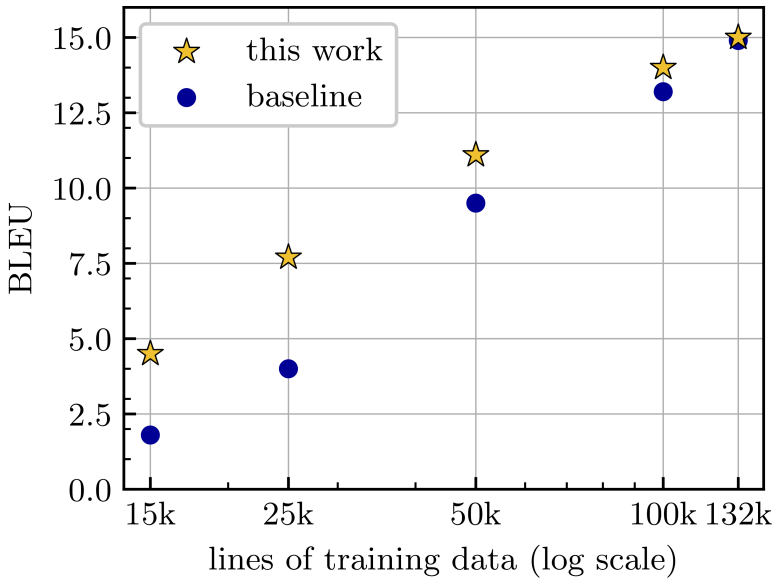


Figure 3.2: Bengali-English data ablation. Improvements of 2.7, 3.7, 1.6, and 0.8 BLEU at the 15k, 25k, 50k, and 100k subsets are statistically significant.

Table 3.2 shows the results for back-translation, our work, and the combination of both.[11] Adding our method to back-translation improves results by an additional 0.5 to 5.7 BLEU.[12]

dataset	GlobalVoices										MATERIAL	
* → en	hu	id	cs	sr	ca	sw	nl	pl	mk	ar	sw	tl
train lines	8k	8k	11k	14k	15k	24k	32k	40k	44k	47k	19k	46k
baseline	2.3	5.3	3.4	11.8	16.0	17.9	22.2	16.0	27.0	12.7	37.8	32.5
baseline w/ back-translation	2.8	7.1	4.6	17.6	20.1	20.7	26.9	19.3	29.1	16.0	38.8	33.0
this work	**5.4**	12.3	**6.6**	16.1	20.0	20.5	24.8	18.0	28.2	14.9	39.0	**33.7**
this work w/ back-translation	4.9	**12.8**	**6.6**	**19.6**	**23.4**	**23.0**	**27.5**	**20.2**	**29.7**	**16.8**	**39.3**	**33.7**

Table 3.2: Comparison between back-translation and this work. We **bold** the best BLEU score on the test set, as well as any result where the difference from it is not statistically significant at the 95% confidence level.

[11] We use a 1:1 ratio between the parallel corpus of and the synthetic back-translated parallel corpus. We use newscrawl2016 (`data.statmt.org/news-crawl`) as monolingual text. When combining with our work, we run our method on both the original and back-translation data.

[12] All statistically significant at the 95% confidence level.

For all language pairs, the best translation quality is achieved by our method combined with back-translation, or our method alone. For 9 of 12 corpora, back-translation and our proposed method are complementary, with improvements of 1.2 to 7.8 BLEU[12] over the baseline when combining the two. For cs-en and tl-en, adding back-translation to our method does not change translation quality as measured by BLEU. In the lowest-resource setting (hu-en) our method alone outperforms the baseline by 3.1 BLEU, but adding back-translation reduces the improvement by 0.5 BLEU.

3.5.3 Method Ablation

In Table 3.3 we analyze the contributions of each part of our proposed method. We compare four conditions to the baseline:[13]

1. paraphrasing the reference, without sampling or the distribution in the loss;[14]
2. sampling from the paraphraser, without the distribution in the loss;
3. using the distribution in the training objective, without sampling the paraphrase; and
4. the proposed method.

[13] All use settings from Section 3.3.2: we use the original reference with $\mathcal{L}_{\text{NLL}}$ with $1 - p = 0.5$ probability, and when sampling we sample from the top $w = 100$ tokens.

[14] This is equivalent to $\mathcal{L}_{\text{NLL}}$ using a paraphrase generated with greedy-search as the reference, see Section 3.5.4.

		dataset	GlobalVoices										MATERIAL	
dist. loss	paraphrase sampling	* → en train lines	hu 8k	id 8k	cs 11k	sr 14k	ca 15k	sw 24k	nl 32k	pl 40k	mk 44k	ar 47k	sw 19k	tl 46k
✗	n/a	baseline	2.3	5.3	3.4	11.8	16.0	17.9	22.2	16.0	27.0	12.7	37.8	32.5
✗	✗	(1)	2.9	8.8	4.6	14.5	17.8	19.2	23.4	17.6	27.0	14.2	35.7	29.9
✗	✓	(2)	5.1	11.6	**6.5**	15.6	**19.7**	**20.2**	24.4	**18.1**	**27.9**	**15.0**	38.1	32.0
✓	✗	(3)	4.0	10.5	**6.5**	15.2	18.8	19.8	23.9	**18.0**	**27.6**	14.4	37.6	31.6
✓	✓	(4) this work	**5.4**	**12.3**	**6.6**	**16.1**	**20.0**	**20.5**	**24.8**	**18.0**	**28.2**	**14.9**	**39.0**	**33.7**

Table 3.3: We compare four conditions to the baseline: (1) paraphrasing the reference, without sampling or the distribution in the loss; (2) sampling from the paraphraser in the training objective, without the distribution; (3) using the distribution in the training objective, without sampling; and (4) the proposed method. We **bold** the best test set BLEU score, and others where the difference is not statistically significant at the 95% confidence level.

We find that sampling is particularly important to the success of our method; removing it significantly degrades translation quality in all but 3 language pairs. Since we sample a paraphrase each batch, this exposes the model to a wide variety of different paraphrases. Using the distribution in the loss function is also beneficial, particularly for the lower resource settings and in the MATERIAL corpora.

3.5.4 Sequence-Level Paraphrastic Data Augmentation

As a contrastive experiment, we use the paraphraser to generate additional target-side data for use in data augmentation. For each target sentence (y) in the training data, we generate a paraphrase (y'). We then concatenate the original source-target pairs (x, y) with the paraphrased pairs (x, y') and perform standard

$\mathcal{L}_{\text{NLL}}$ training. We consider 3 methods for generating paraphrases: beam search (beam of 5), greedy search, sampling (top-100 sampling). Greedy search tends to work best: see Table 3.4. It improves over the baseline for the 10 Global Voices datasets, but not for the two MATERIAL ones. Overall, our proposed method is more effective than this contrastive method. We hypothesize this is due to the wider variety of paraphrases SMRT introduces by sampling and training toward the full distribution from the paraphraser. However, sequence level paraphrastic data augmentation may still be useful in constrained situations where a black-box MT system is used and only the data can be modified.[15]

dataset	GlobalVoices										MATERIAL	
* → en	hu	id	cs	sr	ca	sw	nl	pl	mk	ar	sw	tl
train lines	8k	8k	11k	14k	15k	24k	32k	40k	44k	47k	19k	46k
baseline	2.3	5.3	3.4	11.8	16.0	17.9	22.2	16.0	27.0	12.7	37.8	32.5
beam-search paraphrase	2.6	8.7	4.7	13.5	16.3	18.4	22.6	16.6	26.6	12.2	35.9	29.4
greedy paraphrase	3.2	9.4	4.6	14.8	18.3	19.6	24.4	**18.0**	**27.5**	**14.7**	35.8	30.3
sampled paraphrase	2.8	8.0	5.1	13.9	16.8	19.5	23.9	17.6	**27.6**	14.2	37.2	31.6
this work	**5.4**	**12.3**	**6.6**	**16.1**	**20.0**	**20.5**	**24.8**	**18.0**	**28.2**	**14.9**	**39.0**	**33.7**

Table 3.4: We compare three ways of generating paraphrases for preprocessed data augmentation: beam search, greedy search, and sampling. We **bold** the best BLEU score on the test set, as well as any result where the difference from it is not statistically significant at the 95% confidence level.

[15]Given such a constrained setting multiple different sampled translations could be generated and paired with the original source, while retaining a 1-to-1 ratio of original to paraphrased text, to mimic the sampling portion of our method, and increase the coverage provided by the paraphrases.

3.6 Related Work

3.6.1 Knowledge Distillation

Our proposed objective is similarly structured to word-level knowledge distillation (Hinton et al., 2015; Kim and Rush, 2016, for a more detailed discussion see Chapter 2), where a student model is trained to match the output distribution of a teacher model. Paraphrasing as preprocessed data augmentation, as discussed in Section 3.5.4, is similarly analogous to sequence-level knowledge distillation (Kim and Rush, 2016).

In typical knowledge distillation both the student and teacher models are translation models trained on the same data, have the same input and output languages, and use the original reference for the previous token. In contrast, our teacher model is a paraphraser, which takes as input the original reference sentence (in the target language), with the sampled paraphrase as the previous token. Knowledge distillation is usually used to train smaller models and does not typically incorporate additional data sources, though it has been used for domain adaptation (Dakwale and Monz, 2017; Khayrallah et al., 2018a).

3.6.2 Paraphrasing for Machine Translation

In Hu et al. (2019a), we present case studies on paraphrastic data augmentation for NLP tasks, including neural machine translation. We use sequence-level augmentation

with heuristic constraints on the model's output. SMRT differs in that we train toward the paraphraser *distribution*, and we *sample* from the distribution rather than using heuristics.

Wieting et al. (2019a) used a paraphrastic-similarity metric for minimum risk training (MRT; Shen et al., 2016) in NMT. They note MRT is slow, and, following prior work, use it for fine-tuning after NLL training.

Paraphrasing was also used for statistical MT, including: *source-side*[16] phrase table augmentation (Callison-Burch et al., 2006; Marton et al., 2009), and generation of additional references for tuning (Madnani et al., 2007; Madnani et al., 2008).

3.6.3 Data Augmentation in NMT

Back-translation translates target-language monolingual text to create synthetic source sentences (Sennrich et al., 2016b). Similar to SMRT, it is using an external model to generate additional data. However, back-translation needs a reverse translation model for each *language pair*. In contrast, we need a paraphraser for each *target language*. Zhou et al. (2019) found back-translation is harmful in some low-resource settings, but a strong paraphraser can be trained as long as the target language is sufficiently high resource.

Fadaee et al. (2017) insert rare words in novel contexts in the existing parallel

[16]We were initially inspired by such work, and considered source-side paraphrastic augmentation. In initial experiments, as well as in Hu et al. (2019a), we found that target-side augmentation was more effective.

corpus, using automatic word alignment and a language model. RAML (Norouzi et al., 2016) and SwitchOut (Wang et al., 2018b) randomly replace words others from the vocabulary. In contrast to random or targeted word replacement, we generate semantically similar sentential paraphrases.

3.6.4 Label Smoothing

Label smoothing (Szegedy et al., 2016; Pereyra et al., 2017, which we use when training with $\mathcal{L}_{\text{NLL}}$) spreads probability mass over all non-reference tokens equally; $\mathcal{L}_{\text{SMRT}}$ places higher probability on semantically plausible tokens.

3.6.5 Language Model Integration in NMT

Similar to using a language model in neural machine translation, SMRT incorporates additional target-side data. The paraphraser is conditioned on the full reference, so it can directly replace the reference and captures meaning—not just fluency. Using a language model to rescore an N-best list (Schwenk, 2007; Schwenk, 2012) or interpolating language models (Gülçehre et al., 2015; Gülçehre et al., 2017; Domhan and Hieber, 2017; Stahlberg et al., 2018) only introduce new relative scores; paraphrasing can introduce new target side words.

3.7 Conclusion

We present Simulated Multiple Reference Training (SMRT), which uses transfer learning from a paraphraser to improve translation quality in low-resource settings—by 1.2 to 7.0 BLEU—and is complementary to back-translation.

Neural paraphrasers are rapidly improving (Wieting et al., 2017; Li et al., 2018; Wieting and Gimpel, 2018; Hu et al., 2019a; Hu et al., 2019b; Hu et al., 2019c; Wieting et al., 2019b), and the concurrently released Prism multi-lingual paraphraser (Thompson and Post, 2020a; Thompson and Post, 2020b) has coverage of 39 languages and outperforms prior work in English paraphrasing. As paraphrasing continues to improve and cover more languages, we are optimistic SMRT will provide larger improvements across the board—including for higher-resource MT and for additional target languages beyond English.

Chapter 4

Improving Supervised Domain Adaptation with a Regularized Training Objective

4.1 Introduction

In Chapter 3, we considered the situation where there was an insufficient amount of parallel text in the language pair of interest. In this chapter,[1] we consider the situation where there is a sufficient amount of parallel text in the language pair of interest, but there is an insufficient amount of parallel text in the language pair and *domain* of interest. We focus on the supervised domain adaptation problem, where in addition to a large out-of-domain corpus,[2] we also have a smaller in-domain parallel corpus available for training.[3]

A technique commonly applied in this situation is continued training[4] (Luong and Manning, 2015), where a model is first trained on the out-of-domain corpus, and then that model is used to initialize a new model that is trained on the in-domain corpus.

This simple method leads to empirical improvements on in-domain test sets. However, we hypothesize that some knowledge available in the out-of-domain data—which is not observed in the smaller in-domain data but would be useful at test time—is being forgotten during continued training, due to overfitting. This phenomenon can be viewed as a version of catastrophic forgetting (Goodfellow et al., 2013), a perceptive we explore in Thompson et al. (2019b).

[1] The work described in this chapter was published in Khayrallah et al. (2018a).

[2] This can also be referred to as a 'general domain' corpus. We use 'out-of-domain' in this chapter for consistency with the notation in the published version of this work.

[3] Another challenge in machine translation is the situation where there is no in-domain data available, we do not address that problem in this work.

[4] This is also often referred to as *fine tuning*, we use the term *continued training* to distinguish from the framework of Hinton and Salakhutdinov (2006), which uses supervised learning to fine tune features obtained through unsupervised learning (and for consistency with the notation in the published version of this work).

To address this limitation, we add an additional term to the loss function of the NMT training objective during continued training. In addition to the original term—which minimizes the cross entropy between the model's output distribution and the reference translation—the additional term in the loss function minimizes the cross entropy between the output distribution of the model we are training and the output distribution of the out-of-domain model. This prevents the distribution of words produced from differing too much from the original distribution.

4.2 Method

We focus on the following scenario: we assume there is a model that was trained on a large, general (out-of-domain) corpus in the language pair of interest, and there is a new domain, along with a small in-domain training set, for which we would like to build a model. We begin by initializing the weights of the in-domain model with the weights of the out-of-domain model, and then continue training the new model on the in-domain data, using the modified training objective to prevent the model from differing too much from the original out-of-domain model.

We review the standard NLL training objective and standard continued training, then introduce our proposed objective.

4.2.1 NLL Objective

The standard negative log likelihood (NLL) training objective in NMT, for the i^{th} target word in the reference y is:

$$\mathcal{L}_{\text{NLL}} = -\sum_{v \in \mathcal{V}} \Big[\mathbb{1}\{y_i = v\} \times \log\ p(y_i = v \,|\, x, y_{j<i}) \Big] \tag{4.1}$$

where $\mathcal{V}$ is the vocabulary, $\mathbb{1}\{\cdot\}$ is the indicator function, and $p(y_i = v \,|\, x, y_{j<i})$ is the MT output distribution (conditioned on the source x, and on the previous tokens in the reference $y_{j<i}$). Equation 4.1 computes the cross-entropy between the MT model's distribution and the human gold-standard distribution ($\mathbb{1}\{y_i = v\}$, which is simply a one-hot vector that indicates the correct word).

4.2.2 Continued Training

Continued training is a simple yet effective technique for domain adaptation. It consists of three steps:

1. Train a model until convergence on large out-of-domain parallel corpus using $\mathcal{L}_{\text{NLL}}$ as the training objective.

2. Initialize a new model with the final parameters of Step 1.

3. Train the model from Step 2 until convergence on in-domain parallel corpus, again using $\mathcal{L}_{\text{NLL}}$ as objective.

In other words, continued training initializes an in-domain model training process with parameters from an out-of-domain model. The motivation is that the out-of-domain model provides a reasonable starting point and is better than random initialization.

In our work, we replace $\mathcal{L}_{\text{NLL}}$ in Step 3 by an interpolated regularized objective. All other steps remain the same.

4.2.3 Regularized NMT Objective

We use the output distribution of the trained out-of-domain model to regularize the training of our in-domain model as we perform continued training to adapt to a new domain.

We add an additional regularization (*reg*) term to incorporate information from an auxiliary (*aux*) out-of-domain model in the training objective:

$$\mathcal{L}_{\text{reg}} = -\sum_{v \in \mathcal{V}} \big(\, p_{aux}(y_i = v \,|\, x; y_{j<i}) \times \log \; p(y_i = v \,|\, x; y_{j<i}) \big) \tag{4.2}$$

where p_{aux} is the output distribution from the auxiliary out-of-domain model,[5] and p is the output distribution from the in-domain model being trained.

$\mathcal{L}_{\text{reg}}$ (Equation 4.2) minimizes the cross-entropy between the out-of-domain model distribution $p_{\text{aux}}(y_i = v \mid x; y_{j<i})$ and the in-domain model distribution $p(y_i = v \mid x; y_{j<i})$. We interpolate this with the standard training objective ($\mathcal{L}_{\text{NLL}}$, Equation 4.1) to obtain the final training objective:

$$\mathcal{L} = (1 - \alpha)\, \mathcal{L}_{\text{NLL}} + \alpha\, \mathcal{L}_{\text{reg}} \tag{4.3}$$

4.3 Experiments

4.3.1 Data

We translate from English (en) to German (de) as well as from German to English. For our large, out-of-domain corpus we utilize parallel corpora from WMT2017 (Bojar et al., 2017),[6] which contains data from several sources: Europarl parliamentary proceedings (Koehn, 2005),[7] News Commentary (political and economic news commentary),[8] Common Crawl (web-crawled parallel corpus), and the EU Press Releases.

We use `newstest2015` as the out-of-domain development set and `newstest2016`

[5] The out-of-domain model is fixed while training the in-domain model.
[6] statmt.org/wmt17
[7] statmt.org/europarl
[8] casmacat.eu/corpus/news-commentary.html

as the out-of-domain test set. These consist of professionally translated news articles released by the WMT shared task.

We perform adaptation to two different domains: EMEA (descriptions of medicines) and TED Talks (rehearsed presentations). For EMEA, we use the data split from Koehn and Knowles (2017),[9] which was extracted from OPUS (Tiedemann, 2009; Tiedemann, 2012).[10] For TED, we use the data split from the Multitarget TED Talks Task (MTTT) (Duh, 2018),[11] which was extracted from WIT^3 (Cettolo et al., 2012).[12] Tables 4.1, 4.2, and 4.3 give the number of words and sentences of each of the corpora in the train, dev, and test sets, respectively.

In addition to experiments on the full training sets, we also conduct experiments adapting to each given domain using only the first 2,000 sentences of each in-domain training set to simulate adaptation to a very low-resource domain.

corpus	**de words**	**en words**	**sentences**
EMEA	13,572,552	14,774,808	1,104,752
TED	2,966,837	3,161,544	152,609
WMT	139,449,418	146,569,151	5,919,142

Table 4.1: Tokenized training set sizes.

[9] github.com/khayrallah/domain-adaptation-data
[10] opus.nlpl.eu/EMEA.php
[11] cs.jhu.edu/ kevinduh/a/multitarget-tedtalks
[12] wit3.fbk.eu

corpus	de words	en words	sentences
EMEA	26479	28838	2000
TED	37509	38717	1958
newstest15	44869	47569	2169

Table 4.2: Tokenized development set sizes.

corpus	de words	en words	sentences
EMEA	31737	33884	2000
TED	35516	36857	1982
newstest16	64379	65647	2999

Table 4.3: Tokenized test set sizes.

4.3.2 NMT Settings

Our neural machine translation systems are trained using a modified version of OpenNMT-py (Klein et al., 2017).[13]

We build RNN-based encoder-decoder models with attention (Bahdanau et al., 2015), and use a bidirectional-RNN for the encoder. The encoder and decoder both have 2 layers with LSTM hidden sizes of 1024. Source and target word vectors are of size 500. We apply dropout with 30% probability. We use stochastic gradient descent as the optimizer, with an initial learning rate at 1 and a decay of 0.5. We use a batch size of 64 sentences. We keep the model parameters settings constant for all experiments.

We train byte pair encoding segmentation models (BPE; Sennrich et al., 2016c) on the out-of-domain training corpus. We train separate BPE models for the source and

[13]The code is available: github.com/khayrallah/OpenNMT-py-reg

target language, each with a vocab size of 50,000. We then apply those models to each corpus, including the in-domain ones. This setup allows us to mimic the realistic setting where the computationally-expensive-to-train generic model is trained once, and when there is a new domain that needs translating the existing model is adapted to that domain without retraining on the out-of-domain corpus.

We train our out-of-domain models on the WMT corpora and use the WMT development set (`newstest15`) to select the best epoch as our out-of-domain model. When training our domain specific models, we use the in-domain development set to select the best epoch. When we switch to the in-domain training corpus, we reset the learning rate to 1, with a decay of 0.5, and continue to apply dropout with 30% probability.

4.4 Results

Table 4.4 shows the in-domain and out-of-domain baselines, the improvement provided by continued training, and the added improvement of regularization during continued training on the entire in-domain datasets.[14]

The trends are similar in all four test conditions: Continued training outperforms both baselines, beating the stronger of the two by between 4.0 and 5.3 BLEU points. Our regularization method provides additional improvement over continued training

[14]For the regularized results, α is selected to maximize BLEU on the dev set. See Section 4.5 for more details.

training condition	De-En		En-De	
	EMEA-test	TED-test	EMEA-test	TED-test
out-of-domain (WMT)	30.8	29.8	25.1	25.9
in-domain	43.2	31.4	37.0	25.1
continued-train w/o regularization	48.5	36.4	41.0	30.8
continued-train w/ regularization	49.3 (+0.8)	36.9 (+0.5)	42.5 (+1.5)	30.8 (+0.0)

Table 4.4: BLEU score improvements over continued training. We compare to the out-of-domain baseline and the in-domain baseline. We also compare to continued training without the additional regularization term.

by up to to 1.5 BLEU. There is one setting (En-De TED) where there is no change.

We also repeat the experiment for cases where the in-domain training data is smaller, which corresponds to a more challenging (yet often realistic) domain adaptation scenario. Table 4.5 shows the results of adaptation when only $2,000$ sentences of in-domain parallel text are available. This amount of data is insufficient to train an in-domain NMT model; however, standard continued training is able to improve upon the out-of-domain baseline by 2.2 to 4.9 BLEU. Adding our additional regularization term improves translation quality by an additional 0.2 to 0.9 BLEU.

training condition	De-En		En-De	
	EMEA-test	TED-test	EMEA-test	TED-test
out-of-domain (WMT)	30.8	29.8	25.1	25.9
continued-train w/o regularization	34.3	33.4	30.0	28.1
continued-train w/ regularization	35.2 (+0.9)	33.6 (+0.2)	30.2 (+0.2)	28.4 (+0.3)

Table 4.5: BLEU score improvements over continued training using the $2,000$ sentence subsets as the in-domain corpus. We compare to the out-of-domain baseline and continued training without the additional regularization term.

In both Table 4.4 and Table 4.5, we confirm previous research findings that

continued training is effective, and demonstrate that our regularized objective adds further improvements.

4.5 Analysis

In this section, we perform more detailed analysis of our method. Our research questions are:

1. Is the additional training objective transferring general knowledge to the in-domain model? (Section 4.5.1)
2. What is the impact on translation quality in the original domain? (Section 4.5.2)
3. Why does EMEA show larger improvements? (Section 4.5.3)
4. What value should α be set to? (Section 4.5.4)

4.5.1 Transfer of General-Domain Knowledge

We hypothesize that the regularization term presents knowledge from the out-of-domain model to the continued training model while the model adapts during continued training. This allows the domain-adapted model to retain knowledge from the original (out-of-domain) model that is useful and would otherwise be lost while training continues on the in-domain data, due so the sparsity of the smaller in-domain dataset.

If this is true, using the additional regularization term should improve translation quality of an in-domain model (that does not use continued training), since our technique should transfer general domain knowledge learned from the out-of-domain corpus.

To test this we train an in-domain model from scratch (on only the in-domain data, as opposed to initializing with the general-domain model) using our regularization term. The results are shown in Table 4.6. In this setting, the only out-of-domain information is coming from the additional term in the loss function. Our method provides an improvement of up to 2.3 BLEU over the in-domain model, though in De-En TED translation quality degrades by 0.2 BLEU. While none of these experiments outperform continued training, the large improvements suggest the method is transferring general domain knowledge to the domain specific model.

	De-En		En-De	
training condition	EMEA-test	TED-test	EMEA-test	TED-test
out-of-domain (WMT)	30.8	29.9	25.1	25.9
in-domain	43.2	31.4	37.0	25.1
in-domain w/ regularization	45.5 (+2.3)	31.2 (−0.2)	38.8 (+1.8)	26.0 (+0.9)
continued-train w/o regularization	34.3	33.4	30.0	28.1
continued-train w/ regularization	35.2 (+0.9)	33.6 (+0.2)	30.2 (+0.2)	28.4 (+0.3)

Table 4.6: Analysis of BLEU score improvements without continued training. We compare to the out-of-domain baseline and the in-domain baseline. We show the continued-training results for comparison.

Additionally, these experiments suggest our method could be beneficial in situations where continued training is not an option. For example, the out-of-domain model

might be much larger or perhaps a completely different architecture than the in-domain model; as long as it provides a distribution over the same vocabulary as the in-domain model, it can be used as the auxiliary model in the training objective.

4.5.2 Impact on Original Domain Translation Quality

To examine how well general domain knowledge is retained by the adapted models, we evaluate the domain specific models on a more general domain test set (`newstest2016`),[15] as well as on the other domain's test set (i.e. translation quality of the TED model on the EMEA test set and vice-versa). We report the results for De-En in Table 4.7. In each case, as regularization increases, both general-domain and cross-domain translation quality increase. Continued training for a particular domain harms translation quality on the other domains when compared to the original out-of-domain model.

This suggests that there is some amount of general information about translating between the two languages that is being forgotten by the network during continued training, and the regularization term helps remember it.

[15]Note that this analysis is complicated by the fact that the WMT task is not a single-domain task, since the WMT test set consists of news articles, while the training data includes parliamentary text, political and economic commentary and press releases.

training domain	testset	Baseline		Continued Training (α)			
		in-domain	out-of-domain	0	0.001	0.01	0.1
EMEA	EMEA-dev	49.6	31.4	53.2	53.1	53.4	52.9
	EMEA-test	43.2	30.8	48.5	48.5	49.3	48.1
	newstest2016	5.5	33.8	23.6	23.8	24.1	27.0
TED	TED-dev	27.1	27.1	31.8	31.9	32.2	32.1
	TED-test	27.1	29.8	36.4	36.7	36.9	36.7
	newstest2016	17.0	33.8	30.6	30.9	30.9	31.6

Table 4.7: Analysis of the sensitivity of BLEU scores on the domain-specific sets and `newstest2016` to the interpolation parameter (α) for De-En. Continued training with an $\alpha = 0$ is standard continued training, without regularization. Translation quality of the in-domain test sets is best with an interpolation weight of 0.01 in this language pair, while translation quality of the out-of-domain test sets is better with an interpolation weight of 0.1, the highest value we search over.

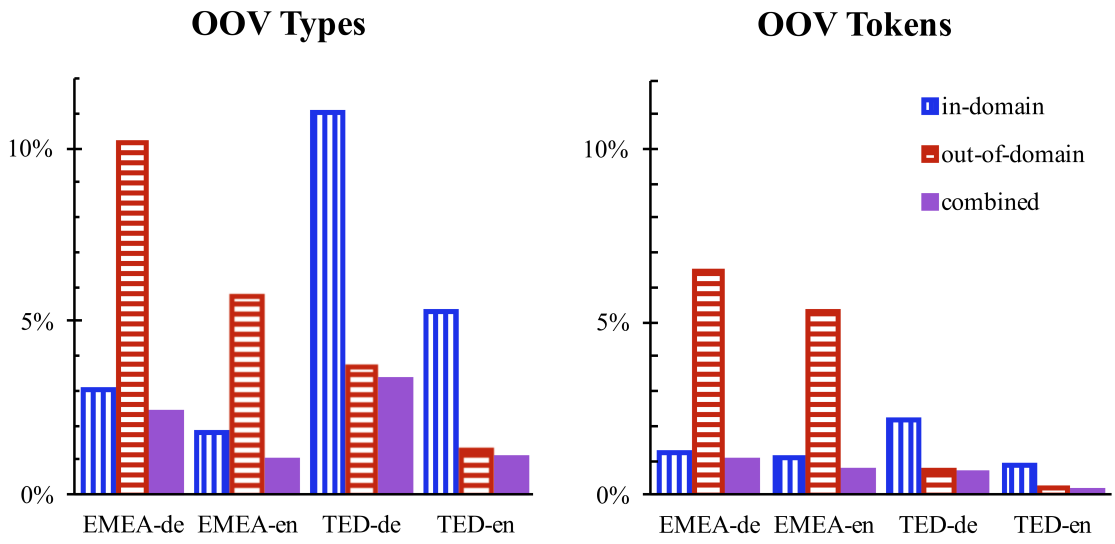
4.5.3 Differences between Domains

Throughout our experiments, we observe larger improvements for EMEA than we do for for TED. For TED, translation quality is similar for both the in-domain and out-of-domain baselines (the in- and out-of-domain baselines are within 1.6 BLEU of each other for TED, whereas for EMEA the in-domain model is over 11 BLEU better in both directions—see Table 4.4 for full results).

We hypothesize that this is because TED is actually similar in domain to our 'out-of-domain' training set. In particular, we suspect that TED talks are similar to parliamentary speech, which are part of the WMT training data—both are oral presentations that cover a variety of topics.

In contrast, EMEA focuses on a single topic (descriptions of medicines) and contains specialized medical terminology throughout.

The out-of-vocabulary rates (OOV) are consistent with this hypothesis (see Tables 4.1a and 4.1b for OOV rates by type and token, respectively). For EMEA, the OOV rate is lower for the in-domain training set compared to the out-of-domain training set while for TED, the opposite is true: the OOV rate is lower for the out-of-domain training set compared to the in-domain training set. This suggests that the EMEA domain has a unique vocabulary that needs to be adapted to, while TED covers a wide variety of topics, and requires a large corpus to cover its vocabulary, and the adaptation problem is more about the style of the corpus.



(a) OOVs by type.

(b) OOVs by token.

Figure 4.1: Percentage of out-of-vocabulary words by (a) *type* and (b) *token*.

This contrast between a very homogeneous domain and a heterogeneous one is typically not made: both are typically described as 'domain adaptation.' However, perhaps future work should approach these problems differently.

4.5.4 Sensitivity to α

We perform a search over α, the interpolation parameter between NLL and our regularization term. We run experiments with α values of 0.001, 0.01, 0.1, and select the best model based on in-domain development set translation quality. Table 4.7 shows the development and test scores when translating to English (the trend is similar translating to German, and is thus not shown here). In general, we see the best in-domain translation quality with α set to 0.01 or 0.1.[16]

4.6 Related Work

Prior work has included the use of similar techniques to solve problems different than ours, as well as different approaches to solve the same problem.

4.6.1 Knowledge Distillation

The added regularization term is formulated in the spirit of knowledge distillation (Hinton et al., 2015; Kim and Rush, 2016, for a more detailed discussion see Chapter 2), where a student model is trained to match the output distribution of a parent model. In word-level knowledge distillation, the student model's output distribution is trained on the same data that the parent model was trained. In contrast, our domain specific model (which replaces the student) is trained with a loss term that encourages it to

[16]It is maybe possible to make further improvements by searching over a more fine-grained range of α values.

match the out-of-domain model (which replaces the parent) on in-domain training data that the out-of-domain model was not trained on.

4.6.2 Regularization Techniques

We draw inspiration from prior works including Yu et al. (2013), which introduces Kullback-Leibler (KL) divergence between the model being trained and an out-of-domain model as a regularization scheme for speaker adaptation. Their work adapts a context dependent deep neural network hidden Markov model (CD-DNN-HMM) using the KL-divergence between the softmax outputs (modeling tied-triphone states) of a network trained on a large, speaker independent (SI) corpus the model being adapted to a specific speaker, initialized with the SI model. Our technique can also be viewed as an extension of label smoothing (Szegedy et al., 2016; Pereyra et al., 2017), where instead of a simple uniform or unigram word distribution, we use the distribution of an auxiliary NMT model.

4.6.3 Continued Training

Since Luong and Manning (2015) introduced continued training in NMT, it has become the de facto standard for domain adaptation. The method has been surprisingly robust, and in-domain improvements have been shown with as few as tens of in-domain training sentences (Miceli Barone et al., 2017).

Despite the success of continued training, several studies have noted that a model trained via continued training tends to significantly underperform the original model on the original domain. Freitag and Al-Onaizan (2016) found that that ensembling an out-of-domain model with a model trained via continued training can significantly reduce the translation quality drop on the original domain compared to the continued training model alone. In contrast, our work focuses on further improving in-domain results.

Chu et al. (2017) present mixed fine-tuning. They begin by training an out-of-domain NMT model but they continue training on a mix of in-domain and out-of-domain data (with the in-domain data oversampled). They also experiment with tagging each sentence with the domain it comes from, allowing a single system to adapt to multiple domains. In contrast, our method does not require further training on (or even access to) the very large general domain dataset while adapting the model to the new domain.

4.6.4 Regularizing Continued Training

Miceli Barone et al. (2017) share our goal of improving in-domain results and compare three methods of regularization to improve continued training: 1) Bayesian dropout 2) L2 regularization, and 3) *tuneout*, which is similar to Bayesian dropout but instead of setting weights to zero, they are set to the value of the out-of-domain model. They report small improvements (≈ 0.3 BLEU) with Bayesian dropout and

L2, but tuneout results are inconsistent and mostly hurt BLEU. In contrast to all three methods, which regularize the weights of the model, our work regularizes only the output distribution and does not directly control the weights.

The work of Dakwale and Monz (2017) is very similar to ours but focuses on retaining out-of-domain translation quality during continued training, instead of in-domain improvements. They perform multi-objective learning with most of the weight (90%) on the auxiliary objective. By contrast, our training emphasizes the in-domain training objective (weighting the auxiliary objective 0.1% to 10%) and we show much larger in-domain improvements.

4.7 Conclusion

In this work, we focus on the scenario where there was sufficient data in the language pair to train a strong model, and we now have a new domain for which we would like a model, but there is a limited amount of training data in the new domain. We add an additional term to the NMT training objective that minimizes the cross-entropy between the model output vocabulary distribution and an auxiliary model's output vocabulary distribution. We begin by initializing with the out-of-domain model, and then continue training on the in-domain data, using the modified training objective to prevent the model from differing too much from the original out-of-domain model. We report improvements of up to 1.5 BLEU over a strong baseline of continued training

when using the full domain adaptation corpora, and up to 0.9 BLEU over continued training in our extremely low resource domain adaptation setting.

In Thompson et al. (2019b), we explore continued training from the perspective of continual learning of highly related tasks, and directly address the degradation observed in out-of-domain translation quality after continued training (as discussed in Section 4.5.2) as an instance of catastrophic forgetting (Goodfellow et al., 2013), by adapting elastic weight consolidation (Kirkpatrick et al., 2017) for continued training of neural machine translation models.

Chapter 5

Analyzing the Impact of Noise on Machine Translation

5.1 Introduction

Even with improved methods we introduced in Chapters 3 and 4 to better leverage limited existing data, the tried-and-true method for improving translation is gathering more data. One approach to complement the expensive task of paying translators to create data to train on is to crawl the web for existing data. This approach is compatible with a variety of data-driven translation approaches. However, as we demonstrate in this chapter,[1] there are challenges with using such web-crawled data, particularly for neural machine translation.

As a motivating example, consider Table 5.1. We add an equally sized noisy web crawled corpus to a high quality German-English training corpus provided by the shared task of the Conference on Machine Translation (WMT).[2] This addition leads to a 1.2 BLEU point increase for the statistical machine translation system, but degrades the neural machine translation system by 9.9 BLEU.

	NMT	**SMT**
WMT17	27.2	24.0
+ noisy corpus	17.3 (–9.9)	25.2 (+1.2)

Table 5.1: Adding noisy web crawled data (raw data from `paracrawl.eu`) to a WMT 2017 German–English statistical system obtains small gains (+1.2 BLEU), a neural system falls apart (–9.9 BLEU).

The maxim *more data is better* that holds true for statistical machine translation

[1] The work described in this chapter was published in Khayrallah and Koehn (2018).

[2] While additional data is of particular interest in low resource language pairs and domains, here we study the impact of noise in a higher resource setting in order to be able to contrast to known clean data, which can be difficult to do in low resource settings where all data is often noisy in some way.

seems to come with more caveats for neural machine translation. The added data cannot be too noisy. In order to reduce the noise, we first seek to understand it: what kind of noise harms neural machine translation models? We explore several types of noise that occur in the web-crawled corpus and assess their impact by adding synthetic noise to an existing parallel corpus. We find that for almost all types of noise, neural machine translation systems are harmed more than statistical machine translation systems. We discovered that one type of noise—copied source language segments—has a catastrophic impact on neural machine translation quality, leading it to learn a copying behavior that it then excessively applies.

5.2 Real-World Noise

What types of noise are prevalent in crawled web data? We manually examined 200 sentence pairs of the Paracrawl corpus and classified them into several error categories. While the results of such a study depend on how crawling and extraction is executed, the results (see Table 5.2) give some indication of what noise to expect.

We classified any pairs of German and English sentences that are not translations of each other as misaligned sentences. These may be caused by any problem in alignment processes (at the document level or the sentence level), or by forcing the alignment of content that is not actually parallel. Such misaligned sentences are the biggest source of error (41%).

Type of Noise	**Amount**
Okay	23%
Misaligned sentences	41%
Third language	3%
Both English	10%
Both German	10%
Untranslated sentences	4%
Short segments (≤2 tokens)	1%
Short segments (3–5 tokens)	5%
Non-linguistic characters	2%

Table 5.2: Types of noise in the raw Paracrawl corpus.

There are three types of wrong language content (totaling 23%): one or both sentences may be in a language different from German and English (3%), both sentences may be German (10%), or both sentences may be English (10%).

4% of sentence pairs are untranslated, i.e., source and target are identical. 2% sentence pairs consist of random byte sequences, only HTML markup, or Javascript. A number of sentence pairs have very short German and/or English sentences, containing at most 2 tokens (1%) or 5 tokens (5%).[3]

Since it is a very subjective value judgment what constitutes disfluent language, we do not classify these as errors. However, consider the sentence pairs in Table 5.3 that we did count as 'okay,' although they contain mostly untranslated names and numbers.

[3]When this work was published, there was concern that short segments (such as glossary entries) might harm the 'language modeling' component of the join neural translation models (specifically RNN models). This work suggested, and further work confirmed, that short segments alone are not the problem, though the quality of those segments matters.

de:	Anonym 2 24.03.2010 um 20:55 314 Kommentare
en:	Anonymous 2 2010-03-24 at 20:55 314 Comments
de:	< < erste < zurück Seite 3 mehr letzte > >
en:	< < first < prev. page 3 next last > >

Table 5.3: Example 'okay' sentences pairs from the paracrawl corpus that might not be ideal for training.

At first glance, some types of noise seem to be easier to automatically identify than others. However, consider, for instance, content in a wrong language. While there are established methods for language identification,[4] these do not work well on a sentence-level basis, especially for lower-resource languages (Caswell et al., 2020), and short sentences (Carter et al., 2013). Or, consider the seemingly obvious problem of untranslated sentences. If they are completely identical, that is easy to spot—although even those may have value, such as the list of country names which are often spelled identical in different languages. There are many degrees of near-identical content of unclear utility.

5.3 Types of Noise

The goal of this work is not to develop methods to detect noise, but rather to ascertain the impact of different types of noise on translation quality when present in parallel data. Our findings informed subsequent work on parallel corpus cleaning (see Section 5.7).

[4] github.com/google/cld3

We now formally define five types of naturally occurring noise and describe how we simulate them.[5] By creating artificial noisy data, we avoid the hard problem of detecting specific types of noise but are still able to study their impact.

5.3.1 Misaligned Sentences

As shown above, a common source of noise in parallel corpora is faulty document or sentence alignment. This results in sentences that are not matched to their translation. Such noise is rare in certain corpora such as Europarl (Koehn, 2005)—where strong clues about debate topics and speaker turns reduce the scale of the task of alignment to paragraphs—but more common in the alignment of less structured web sites. We artificially create misaligned sentence data by randomly shuffling the order of sentences on one side of the original clean parallel training corpus.

5.3.2 Misordered Words

Language may be disfluent in many ways. Disfluency may be the product of machine translation, poor human translation, or heavily specialized language use, such as bullet points in product descriptions (recall also the examples above). We consider one extreme case of disfluent language: sentences from the original corpus where the words are reordered randomly. We do this on the source or target side.

[5] We release the simulated data: data.statmt.org/noise

5.3.3 Wrong Language

A parallel corpus may be polluted by text in a third language, say French in a German–English corpus. This may occur on the source or target side of the parallel corpus. To simulate this, we add French–English (bad source) or German–French (bad target) data to a German–English corpus.

5.3.4 Untranslated Sentences

Especially in parallel corpora crawled from the web, there are often sentences that are untranslated from the source in the target. Examples are navigational elements or copyright notices in the footer. Purportedly multilingual web sites may be only partially translated, while some original text is copied. Again, this may show up on the source or the target side. We take sentences from either the source or target side of the original parallel corpus and simply copy them to the other side.

5.3.5 Short Segments

Sometimes additional data comes in the form of bilingual dictionaries. Can we simply add them as additional sentence pairs, even if they consist of single words or short phrases? We simulate this kind of data by subsampling a parallel corpus to include only sentences of maximum length 2 or 5.

5.4 Experimental Setup

5.4.1 Neural Machine Translation

Our neural machine translation systems are trained using Marian (Junczys-Dowmunt et al., 2018).[6] We build RNN-based encoder-decoder models with attention (Bahdanau et al., 2015). We train Byte-Pair Encoding segmentation models (BPE; Sennrich et al., 2016c) with a vocab size of $50,000$ on both sides of the parallel corpus for each experiment. We apply drop-out with 20% probability on the RNNs, and with 10% probability on the source and target words. We stop training after convergence of cross-entropy on the development set, and we average the 4 highest performing models (as determined by development set BLEU) to use as an ensemble for decoding (checkpoint ensembling). Training of each system takes 2–4 days on a single GPU (GTX 1080ti).

5.4.2 Statistical Machine Translation

Our statistical machine translation systems are trained using Moses (Koehn et al., 2007).[7] We build phrase-based systems using standard features commonly used in recent system submissions to WMT (Haddow et al., 2015; Ding et al., 2016; Ding et al., 2017). We train our systems with the following settings: a maximum

[6] marian-nmt.github.io
[7] statmt.org/moses

sentence length of 80, grow-diag-final-and symmetrization of GIZA++ alignments, an interpolated Kneser-Ney smoothed 5-gram language model with KenLM (Heafield, 2011), hierarchical lexicalized reordering (Galley and Manning, 2008), a lexically-driven 5-gram operation sequence model (OSM; Durrani et al., 2013), sparse domain indicator, phrase length, and count bin features (Blunsom and Osborne, 2008; Chiang et al., 2009), a maximum phrase-length of 5, compact phrase table (Junczys-Dowmunt, 2012), minimum Bayes risk decoding (Kumar and Byrne, 2004), cube pruning (Huang and Chiang, 2007), with a stack-size of 1000 during tuning. We optimize feature function weights with k-best MIRA (Cherry and Foster, 2012).

While we focus on phrase based systems as our SMT paradigm, we note that there are other statistical machine translation approaches such as hierarchical phrase-based models (Chiang, 2007) and syntax-based models (Galley et al., 2004; Galley et al., 2006) that may have better translation quality in certain language pairs and in low resource conditions.

5.4.3 Clean Corpus

In our experiments, we translate from German to English. We use corpora from the shared translation task organized alongside the Conference on Machine Translation (WMT)[8] as clean training data. For our baseline we use: Europarl (Koehn, 2005),[9]

[8] statmt.org/wmt17/
[9] statmt.org/europarl

News Commentary,[10] and the Rapid EU Press Release parallel corpus. The corpus size is about 83 million tokens per language. We use `newstest2015` for tuning SMT systems, `newstest2016` as a development set for NMT systems, and report results on `newstest2017`.

We always train our language model for statistical machine translation on the target side of the parallel corpus for that experiment. Note that we do not add monolingual data to our systems since this would make our study more complex. While using monolingual data for language modeling was standard practice in statistical machine translation, how to use such data for neural models was less obvious at the time of this work.[11]

5.4.4 Noisy Corpora

Here we describe how each specific noisy-corpus was created.[12]

For MISALIGNED SENTENCE and MISORDERED WORD noise, we use the clean corpus (above) and perturb the data. To create UNTRANSLATED SENTENCE noise, we also use the clean corpus and create pairs of identical sentences.

For WRONG LANGUAGE noise, we do not have French–English and German–French data of the same size from the same sources. Hence, we use the EU Bookstore corpus

[10] casmacat.eu/corpus/news-commentary.html

[11] As this dissertation is being completed, back-translation (Sennrich et al., 2016b) is the standard method for monolingual data integration in neural machine translation.

[12] The data is available at data.statmt.org/noise.

(Skadiņš et al., 2014).[13]

The SHORT SEGMENTS are extracted from OPUS corpora (Tiedemann, 2009; Tiedemann, 2012; Lison and Tiedemann, 2016):[14] EMEA (descriptions of medicines),[15] Tanzil (religious text),[16] Open Subtitles 2016,[17] Acquis (legislative text),[18] GNOME (software localization files),[19] KDE (localization files), PHP (technical manual),[20] Ubuntu (localization files),[21] and Open Office.[22] We use only pairs where both the English and German segments are at most 2 or 5 words long. Since this results in small data sets (2 million tokens and 15 million tokens per language, respectively), they are duplicated multiple times.

We also show the results for naturally occurring noisy web data from the raw 2016 ParaCrawl corpus (Bañón et al., 2020).[23]

We sample the noisy corpus in an amount equal to 5%, 10%, 20%, 50%, and 100% of the clean corpus. We then combine the noisy corpus with the clean one. This reflects a realistic situation where there is a clean corpus, and one would like to add additional data that has the potential to be noisy. For each experiment, we use the target side of the parallel corpus to train the SMT language model, including the

[13] opus.nlpl.eu/EUbookshop.php
[14] opus.nlpl.eu
[15] emea.europa.eu
[16] tanzil.net/trans
[17] opensubtitles.org
[18] ec.europa.eu/jrc/en/language-technologies/jrc-acquis
[19] l10n.gnome.org
[20] se.php.net/download-docs
[21] translations.launchpad.net
[22] openoffice.org
[23] We use the deduplicated raw set from paracrawl.eu.

noisy text.

5.5 Impact on Translation Quality

Table 5.4 shows the effect of adding each type of noise to the clean corpus.[24] For some types of noise NMT is harmed more than SMT: MISMATCHED SENTENCES (up to -1.9 for NMT, -0.6 for SMT), MISORDERED WORDS (source) (-1.7 vs. -0.3), WRONG LANGUAGE (target) (-2.2 vs. -0.6).

SHORT SEGMENTS, UNTRANSLATED SOURCE SENTENCES and WRONG SOURCE LANGUAGE have little impact on either SMT or NMT (at most a degradation of -0.7). MISORDERED TARGET WORDS decreases BLEU scores for both SMT and NMT by just over 1 point (100% noise).

The most dramatic difference is UNTRANSLATED TARGET SENTENCE noise. When added at 5% of the original data, it degrades NMT quality by 9.6 BLEU, from 27.2 to 17.6. Adding this noise at 100% of the original data degrades quality by 24.0 BLEU, dropping the score from 27.2 to 3.2. In contrast, the SMT system only drops 2.9 BLEU, from 24.0 to 21.1.

[24] We report case-sensitive detokenized BLEU (Papineni et al., 2002) calculated using `mteval-v13a.pl`.

	5%	10%	20%	50%	100%
Misaligned sentences	26.5 24.0 -0.7 -0.0	26.5 24.0 -0.7 -0.0	26.3 23.9 -0.9 -0.1	26.1 23.9 -1.1 -0.1	25.3 23.4 -1.9 -0.6
Misordered words (source)	26.9 24.0 -0.3 -0.0	26.6 23.6 -0.6 -0.4	26.4 23.9 -0.8 -0.1	26.6 23.6 -0.6 -0.4	25.5 23.7 -1.7 -0.3
Misordered words (target)	27.0 24.0 -0.2 -0.0	26.8 24.0 -0.4 -0.0	26.4 23.4 -0.8 -0.6	26.7 23.2 -0.5 -0.8	26.1 22.9 -1.1 -1.1
Wrong language (French source)	26.9 24.0 -0.3 -0.0	26.8 23.9 -0.4 -0.1	26.8 23.9 -0.4 -0.1	26.8 23.9 -0.4 -0.1	26.8 23.8 -0.4 -0.2
Wrong language (French target)	26.7 24.0 -0.5 -0.0	26.6 23.9 -0.6 -0.1	26.7 23.8 -0.5 -0.2	26.2 23.5 -1.0 -0.5	25.0 23.4 -2.2 -0.6
Untranslated (English source)	27.2 23.9 -0.0 -0.1	27.0 23.9 -0.2 -0.1	26.7 23.6 -0.5 -0.4	26.8 23.7 -0.4 -0.3	26.9 23.5 -0.3 -0.5
Untranslated (German target)	17.6 23.8 -9.8 -0.2	11.2 23.9 -16.0 -0.1	5.6 23.8 -21.6 -0.2	3.2 23.4 -24.0 -0.6	3.2 21.1 -24.0 -2.9
Short segments (max 2)	27.1 24.1 -0.1 +0.1	26.5 23.9 -0.7 -0.1	26.7 23.8 -0.5 -0.2		
Short segments (max 5)	27.8 24.2 +0.6 +0.2	27.6 24.5 +0.4 +0.5	28.0 24.5 +0.8 +0.5	26.6 24.2 -0.6 +0.2	
Raw crawl data	27.4 24.2 +0.2 +0.2	26.6 24.2 -0.6 +0.2	24.7 24.4 -2.5 +0.4	20.9 24.8 -6.3 +0.8	17.3 25.2 -9.9 +1.2

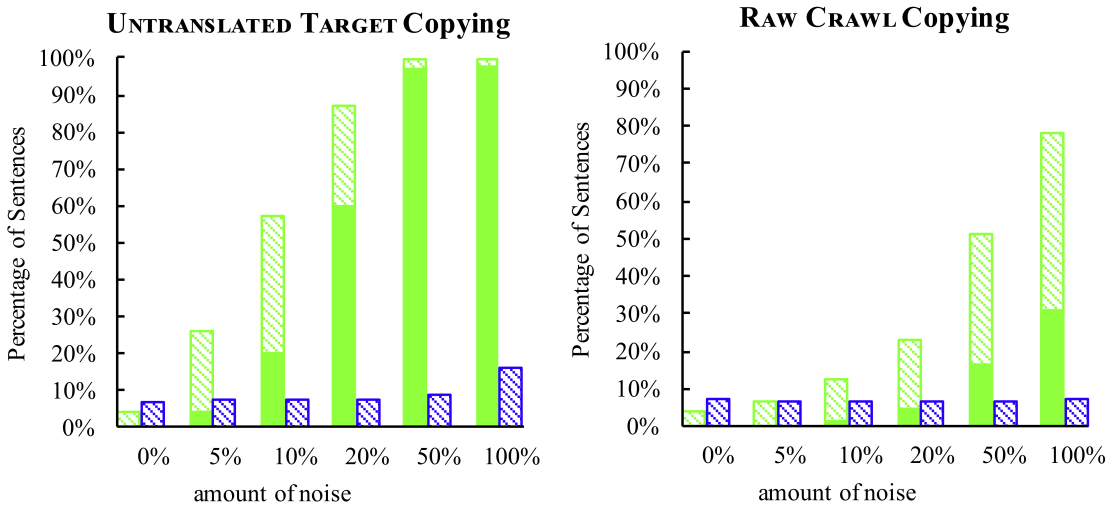
Table 5.4: Results from adding different amounts of noise (ratio of original clean corpus) for various types of noise in German-English Translation. Generally neural machine translation (left green bars) is harmed more than statistical machine translation (right blue bars). The worst type of noise are segments in the source language copied untranslated into the target language.

5.5.1 Copied Output

Since the noise type where the target side is a copy of the source has such a big impact, we examine the system output in more detail.

We report the percent of sentences in the evaluation set that are identical to the source for the UNTRANSLATED TARGET SENTENCE and RAW CRAWL data in Figure 5.1 (solid bars). The SMT systems output 0 or 1 sentences that are exact copies. However, with just 20% of the UNTRANSLATED TARGET SENTENCE noise (which corresponds to 10% of the total training data being noisy, since it is combined with clean data), 60% of the NMT output sentences are identical to the source. This suggests that the NMT systems learn to copy, which may be useful for named entities. However, with even a small amount of this data it is doing far more harm than good.

Figure 5.1 also shows the percent of sentences that have a worse TER score against the reference than against the source (shaded bars). This means that it would take fewer edits to transform the sentence into the source sequence than it would take to transform it into the target sequence. When just 10% UNTRANSLATED TARGET SENTENCE data is added, 57% of the sentences are more similar to the source than to the reference, indicating partial copying. This suggests that the NMT system is overfitting on the copied portion of the training corpus. This is supported by Figure 5.2, which shows the learning curve on the development set for the UNTRANSLATED TARGET SENTENCE noise setup. The translation quality for the systems trained on noisy corpora begin to improve, before over-fitting to the copy portion of the training set.



(a) Untranslated (target) (b) raw crawl

Figure 5.1: Copied sentences in the Untranslated (target) and raw crawl experiments. NMT is the left green bars, SMT is the right blue bars. Sentences that are exact matches to the source are the solid bars, sentences that are more similar to the source than the target are the shaded bars.

Note that while we plot the BLEU score on the development set with beam search, the system is optimizing cross-entropy given a perfect prefix.

Though 7.4% of the sentences in the raw crawl data was exact copied sentences (compared to 1.7% of sentences in the clean data) we find that when using equal amounts of raw-crawled and clean data (the far right column in Table 5.4), 31% of the output sentences were exact copies, and 18% of the remaining sentences were more similar to the source than the reference.

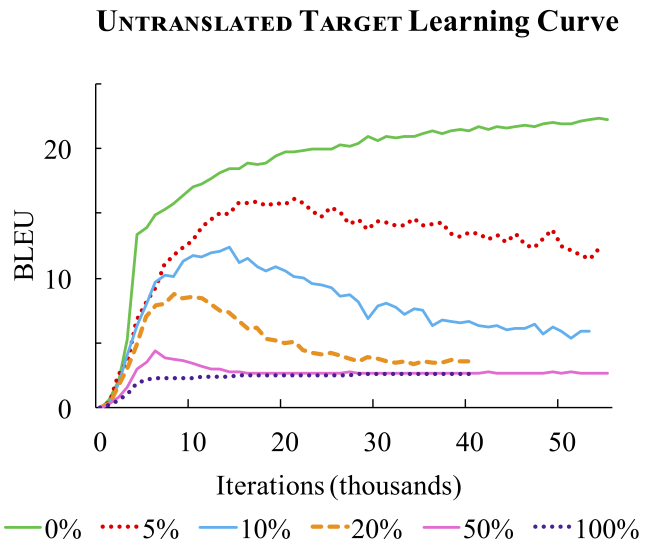


Figure 5.2: Learning curves for the NMT UNTRANSLATED TARGET SENTENCE experiments.

5.5.2 Incorrect Language Output

We performed a manual analysis of the neural machine translation experiments where a German–French corpus is added to a German–English corpus (WRONG TARGET LANGUAGE). For each of the noise levels, Table 5.5 shows the percentage of NMT output sentences in French, out of a total of 3004. Most NMT output sentences were either entirely French or English, with the exception of a few mis-translated cognates (e.g.: 'façade', 'accessibilité').

In the SMT experiment with 100% noisy data added, there are a couple of French words in mostly English sentences. These are much less frequent than unknown German words passed through. Only 1 of the 3004 sentences is mostly French.

At first glance, it is surprising that such a small percentage of the output sentences

Amount of Noise in Training	Amount of French in Output
5%	0.2%
10%	0.6%
20%	1.7%
50%	3.3%
100%	6.7%

Table 5.5: Percentage of the 3004 sentences in the test set that were translated to French when different amounts of WRONG LANGUAGE (FRENCH TARGET) noise (ratio of original clean corpus).

were French, since up to half of the target data in training was in French. We attribute this to the domain of the added data differing from the test data. We essentially had a multi-task model. Source sentences in the test set are more similar to the domain-relevant clean parallel training corpus than the domain-divergent noise corpus. Therefore, the model is able to determine which language it should be producing based on the domain of the input sentence. While we observe a drop in quality (2.2 BLEU when half the target training data was in French) only 6.7% of the total output sentences were in French.

Aharoni and Goldberg (2020) found that neural language models are able to learn sentence representations that cluster sentences according to domain, without domain supervision. This suggests that our translation model may have also been able to learn a domain classification model, based on the input sentences.

5.6 Related Work

There is a robust body of work on filtering out noise in parallel data. For example: Taghipour et al. (2011) use an outlier detection algorithm to filter a parallel corpus; Xu and Koehn (2017) generate synthetic noisy data (inadequate and non-fluent translations) and use this data to train a classifier to identify good sentence pairs from a noisy corpus; and Cui et al. (2013) use a graph-based random walk algorithm and extract phrase pair scores to weight the phrase translation probabilities to bias towards more trustworthy ones.

Most of this work was done in the context of statistical machine translation, but more recent work (Carpuat et al., 2017) targets neural models. That work focuses on identifying semantic differences in translation pairs using cross-lingual textual entailment and additional length-based features, and demonstrates that removing such sentences improves neural machine translation quality.

As Rarrick et al. (2011) point out, one problem of parallel corpora extracted from the web is translations that have been created by machine translation. Venugopal et al. (2011) propose a method to watermark the output of machine translation systems to aid this distinction. Belz and Kow (2011) report that rule-based machine translation output can be detected due to certain word choices, and statistical machine translation output due to lack of reordering.

In 2016, shared tasks were organized on document alignment (Buck and Koehn,

2016), and on sentence pair filtering.[25] The latter was in the context of cleaning translation memories which tend to be cleaner that the data collected from web crawls.

Belinkov and Bisk (2018), Anastasopoulos et al. (2019), and Anastasopoulos (2019) investigate noise in neural machine translation, but they focus on creating systems that can *translate* the kinds of orthographic errors (typos, misspellings, etc.) that humans often produce and can comprehend. In contrast, we address noisy *training* data and focus on types of noise occurring in web-crawled corpora.

There is a rich literature on data selection which aims at sub-sampling parallel data relevant for a task-specific machine translation system (Axelrod et al., 2011). Wees et al. (2017) find that the existing data selection methods developed for statistical machine translation are less effective for neural machine translation. This is different from our goals of handling noise since those methods tend to discard perfectly fine sentence pairs (e.g., software manuals) that are just not relevant for the targeted domain (e.g., social media). Our work is focused on noise that is harmful for all domains.

Other work has also considered copying in NMT. Currey et al. (2017) add copied data and back-translated data to a clean parallel corpus. They report improvements on English↔Romanian when adding as much back-translated and copied data as they have parallel (1:1:1 ratio). For English↔Turkish and English↔German, they add twice as much back translated and copied data as parallel data (1:2:2 ratio), and

[25]NLP4TM 2016 shared task: rgcl.wlv.ac.uk/nlp4tm2016/shared-task

report improvements on English↔Turkish but not on English↔German. However, their English↔German systems trained with the copied corpus did not perform worse than baseline systems.

In work contemporaneous to ours, Ott et al. (2018) found that while copied training sentences[26] represent less than 2.0% of their ('clean') training data (WMT 14 English↔German and English↔French), copies are over-represented in the output of beam search.[27] Using a subset of training data from WMT 17, they replace a subset of the true translations with a copy of the input. They analyze varying amounts of copied noise, and a variety of beam sizes. Larger beams are more affected by this kind of noise; however, for all beam sizes quality degrades completely with 50% copied sentences.[28]

5.7 Impact on Subsequent Work

These findings informed future work on corpus cleaning. As a follow-up to this work, we organized a shared task on filtering web-crawled data (Koehn et al., 2018). In this task, participants were given access to clean parallel data and web-crawled data and asked to identify the clean portion of web crawled data. Participants were judged by how well MT models trained on only their filtered web crawled data performed.

[26]That work defines a copying as any sentence pair where intersection over the union of unigrams (excluding punctuation and numbers) is at least 50%.

[27]They report the following copy rates for various beam sizes on en↔fr: 2.6% (beam=1), 2.9% (beam=5), 3.2% (beam=10) and 3.5% (beam=20)

[28]See Figure 3 in Ott et al. (2018).

This task drew 18 submissions. While a variety of methods were used, the majority included: (1) pre-filtering rules, (2) scoring functions for sentence pairs, and some included (3) a classifier that learned weights for feature functions. Many of the pre-filtering rules reflect the work in this chapter, such as removing sentence pairs of vastly different lengths (which suggests they are not aligned sentences), removing sentence pairs that are too similar (which indicates copy noise), and removing sentences where the language identifier does not detect the required language.

After prefiltering, scoring functions were applied. These include as n-gram or neural language models on clean data, language models trained on the provided raw data as contrast, neural translation models and bag-of-words lexical translation probabilities. The winning submission on Dual Conditional Cross-Entropy Filtering (Junczys-Dowmunt, 2018) set a new state-of-the-art in data filtering.

In subsequent years (Koehn et al., 2019; Koehn et al., 2020), the shared task focused on lower resource settings. Multilingual sentence embeddings (LASER; Artetxe and Schwenk, 2019) were applied for this task (Chaudhary et al., 2019), and performed well, particularly in lower resource settings. This reflects a growing trend in natural language processing as whole towards transfer learning across languages and tasks. Such methods can be particularly important in low resource settings, where there may be insufficient parallel data to train initial models for use in, for example, Dual Conditional Cross-Entropy Filtering.

There was also work to extend dual conditional cross entropy filtering to the case

where no parallel data is available (Axelrod et al., 2019).

For the 2020 iteration of the task, there was more focus on modeling this as a classification task rather than simply ranking. This required both clean and noisy examples to train the classifier. Since the shared task provided some clean training data, participants used that and also corrupted sentences for negative training examples. Some corrupted sentences address challenges highlighted in this chapter, such as mismatched sentences, truncated sentences and sentences with swapped word order (Esplà-Gomis et al., 2020; Açarçiçek et al., 2020; ElNokrashy et al., 2020; Xu et al., 2020)

5.8 Conclusion

We described five types of noise in parallel data, motivated by a manual analysis of raw web crawl data. We found that neural machine translation is less robust to many types of noise than statistical machine translation.

In general, systematic noise has a larger negative impact than random noise. Additionally, noise on the target side tends to be more problematic. Certain types of source noise can distinguished from the 'clean' text, allowing the model to learn a kind of multi-task model. However, if the distinction is on the target side, it cannot be learned in a way that can be identified by the model during inference, when only the source sentence is available.

In particular, copied data—where the target side of the training data is identical to the source—is problematic because the model learns to copy at a much higher rate than copying occurs in the training data. We observed a similar effect of copying being over represented in the output of the experiments trained on raw web crawled data, suggesting careful consideration of overlap between source and target training data is necessary.

While we focus on RNN-based models with attention as our NMT architecture, we note that different architectures have been proposed, including based on convolutional neural networks (Kalchbrenner and Blunsom, 2013; Gehring et al., 2017) and the self-attention based Transformer model (Vaswani et al., 2017).

In this study, we focused on a relatively high resource setting, in order to be able to do controlled experiments in comparison to known clean data. Finding additional data is particularly important for low resource language pairs and domains, however in such settings all of the data available might be noisy (Caswell et al., 2021). This makes it even more crucial to think carefully about data quality and potential filtering approaches.

Chapter 6

Conclusion

6.1 Summary

Despite neural machine translation's increased quality and prevalence, data quality and quantity remain challenges in machine translation.

Limited quantities of such data are available for most language pairs, leading to a *low resource* problem. Even when training data is available in the desired language pair, it is frequently formal speech or news—leading to a *domain* mismatch when models are used to translate a different type of data from most of what they were trained on. Neural machine translation currently performs poorly in domain adaptation and low resource settings (Koehn and Knowles, 2017; Sennrich and Zhang, 2019). An obvious approach when faced with a lack of data is to go get more data. This is often the best way to improve translation quality. However, it is not always feasible to produce additional human translations. In such a case, an option may be to crawl the web for additional training data. However, such data can be very noisy and harm machine translation quality— particularly neural machine translation quality.

This dissertation addresses these three specific data challenges in machine translation:

1. We present a method for transfer learning from a paraphraser to overcome data sparsity in low resource settings Chapter 3.

2. We present a method for improving domain adaptation translation quality, when sufficient data is available in the language pair of interest, but not in the domain

of interest Chapter 4.

3. We consider web-crawls as a method for acquiring more data, and find that such data can harm machine translation quality if not carefully filtered Chapter 5.

6.2 Future Work

This dissertation took steps towards improving translation quality in difficult data settings in order to improve information access and communication for users of all languages. However, there remains future work to be done towards that goal, including:

6.2.1 Revisiting the Impact of Noisy Training Data on NMT

In Chapter 5 we investigated the impact of different types of artificially created noisy training data on NMT, in order to motivate future work on parallel corpus filtering. This work spurred further research in the area, but there were some limitations of the study that merit re-exploration, including:

- Our work used RNNs, but there have been several advances in NMT training in the past three years. An interesting line of work would be to consider more recent neural machine translations architectures as well.

- Our work did not consider any data augmentation or pretraining. It would be interesting to investigate the impact of noisy data on the pretrained models themselves, as well as when performing adaptation. It would also be impactful to understand how data augmentation interacts with noisy data.

- Our work considered the impact of web-crawled noisy data. There are other forms of noise in machine translation data. This includes sampled back-translation data (Edunov et al., 2018), and sampled paraphrase data (Chapter 3; Khayrallah et al., 2020a). These types of data have been demonstrated to improve translation quality, despite adding some noise. A better understanding of the impact of different types of noise may help with improved sampling techniques for data augmentation.

- There was a domain shift in the data added for one of the noise types (Wrong Language) that may have confounded some of the results, this should be corrected in any future work.

- We only considered translating from German to English. This is a relatively high resource pair, with somewhat similar languages. We began we clean corpora and added potentially noisy data. Exploring noise in lower resource settings with more dissimilar languages would be beneficial. This may pose a bit of a challenge, since often all available data in low resource settings is noisy, but would be a realistic use-case (Caswell et al., 2021).

6.2.2 Learning to Learn from Diverse Data

In certain situations there is a sudden need for translation after a disaster, such as the 2010 Haitian Earthquake, where an MT model was requested by first responders to be able to translate text messages sent to a helpline. An MT model was built from scratch in under a week (Lewis, 2010).

This dissertation explored how to leverage different forms of non-standard data to overcome training data sparsity, focusing on a targeted type of data for each identified problem. However, such approaches require some level of specialization for each language pair and domain, depending on the data that might be available, and therefore requires a human expert to experiment with different methods to decide what will work best. That is impractical to scale to all language pairs in the world, and may not always be fast enough.

There is beginning to be some work on learning how to learn from diverse data sets (e.g., Wees et al., 2017; Wang et al., 2018a; Kumar et al., 2019; Wang et al., 2020; Kumar et al., 2021a; Kumar et al., 2021b), though improved computational efficiency is crucial.

6.2.3 Multilingual NLP

Improving communication is about more than just translation. All natural language processing (NLP) tools we build should also serve everyone. Beyond machine

translation, there are other (multilingual) NLP problems that have difficult data settings. Adapting and expanding the techniques such as those developed for difficult data settings in machine translation to other NLP systems have a high potential for impact. Additionally, studies on the effect of noisy data on other NLP tasks, and follow up work on how to mitigate the effect also have a high potential for impact. This applies to both web-crawled data, and somewhat curated data, which may be noisy as well.

6.3 Closing Remarks

In this dissertation, we consider three data challenges that affect machine translation quality, and reconsider what and how different types of available data can be used to improve machine translation.

In recent years, with the move toward end-to-end neural models, there has been a trend toward abstracting many different NLP tasks as sequence to sequence tasks, and focusing on the modeling without an equivalently thorough focus on the data.

While much can be learned from vision, speech, and other NLP tasks and it remains important to learn from adjacent fields, it is crucial to consider the subtleties in different tasks, and to always carefully consider the data available, and also additional data that could be leveraged. *Careful* integration of additional data in an intelligent way can often have a high impact.

Bibliography

Haluk Açarçiçek, Talha Çolakoğlu, Pınar Ece Aktan Hatipoğlu, Chong Hsuan Huang, and Wei Peng (2020). "Filtering Noisy Parallel Corpus using Transformers with Proxy Task Learning". *Proceedings of the Fifth Conference on Machine Translation*. Online: Association for Computational Linguistics, pages 940–946. URL: https://www.aclweb.org/anthology/2020.wmt-1.105 (cited on page 95).

Douglas Adams and Geoffrey Perkins (1985). *The hitch-hiker's guide to the galaxy: the original radio scripts*. English. OCLC: 1244793154. London: Pan. URL: https://archive.org/details/hitchhikersguide0000adam_f5g9 (cited on page 11).

Roee Aharoni and Yoav Goldberg (2020). "Unsupervised Domain Clusters in Pretrained Language Models". *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pages 7747–7763. DOI: 10.18653/v1/2020.acl-main.692. URL: https://www.aclweb.org/anthology/2020.acl-main.692 (cited on page 90).

Antonios Anastasopoulos (2019). "An Analysis of Source-Side Grammatical Errors in NMT". *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing*

and Interpreting Neural Networks for NLP. Florence, Italy: Association for Computational Linguistics, pages 213–223. DOI: 10.18653/v1/W19-4822. URL: https://www.aclweb.org/anthology/W19-4822 (cited on page 92).

Antonios Anastasopoulos, Alison Lui, Toan Q. Nguyen, and David Chiang (2019). "Neural Machine Translation of Text from Non-Native Speakers". *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pages 3070–3080. DOI: 10.18653/v1/N19-1311. URL: https://www.aclweb.org/anthology/N19-1311 (cited on page 92).

Mikel Artetxe and Holger Schwenk (2019). "Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond". *Transactions of the Association for Computational Linguistics* 7, pages 597–610. ISSN: 2307-387X. DOI: 10.1162/tacl_a_00288. eprint: https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00288/1879548/tacl_a_00288.pdf. URL: https://doi.org/10.1162/tacl_a_00288 (cited on page 94).

Amittai Axelrod, Xiaodong He, and Jianfeng Gao (2011). "Domain Adaptation via Pseudo In-Domain Data Selection". *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Edinburgh, Scotland, UK.: Association for Computational Linguistics, pages 355–362. URL: https://www.aclweb.org/anthology/D11-1033 (cited on page 92).

BIBLIOGRAPHY

Amittai Axelrod, Anish Kumar, and Steve Sloto (2019). "Dual Monolingual Cross-Entropy Delta Filtering of Noisy Parallel Data". *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*. Florence, Italy: Association for Computational Linguistics, pages 245–251. DOI: 10.18653/v1/W19-5433. URL: https://www.aclweb.org/anthology/W19-5433 (cited on page 95).

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio (2015). "Neural Machine Translation by Jointly Learning to Align and Translate". Proceedings of the International Conference on Learning Representations (ICLR). URL: http://arxiv.org/pdf/1409.0473v6.pdf (cited on pages 15, 61, 81).

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza (2020). "ParaCrawl: Web-Scale Acquisition of Parallel Corpora". *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pages 4555–4567. DOI: 10.18653/v1/2020.acl-main.417. URL: https://www.aclweb.org/anthology/2020.acl-main.417 (cited on pages 34, 84).

Yonatan Belinkov and Yonatan Bisk (2018). "Synthetic and Natural Noise Both Break Neural Machine Translation". *International Conference on Learning*

Representations. URL: https://openreview.net/forum?id=BJ8vJebC- (cited on page 92).

Anja Belz and Eric Kow (2011). "Unsupervised Alignment of Comparable Data and Text Resources". *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*. Portland, Oregon: Association for Computational Linguistics, pages 102–109. URL: https://www.aclweb.org/anthology/W11-1214 (cited on page 91).

Adrian Benton, Huda Khayrallah, Biman Gujral, Dee Ann Reisinger, Sheng Zhang, and Raman Arora (2019). "Deep Generalized Canonical Correlation Analysis". *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*. Florence, Italy: Association for Computational Linguistics, pages 1–6. DOI: 10.18653/v1/W19-4301. URL: https://www.aclweb.org/anthology/W19-4301 (cited on page 8).

Phil Blunsom and Miles Osborne (2008). "Probabilistic Inference for Machine Translation". *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Honolulu, Hawaii: Association for Computational Linguistics, pages 215–223. URL: https://www.aclweb.org/anthology/D08-1023 (cited on page 82).

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi (2015). "Findings

of the 2015 Workshop on Statistical Machine Translation". *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Lisbon, Portugal: Association for Computational Linguistics, pages 1–46. DOI: 10.18653/v1/W15-3001. URL: https://www.aclweb.org/anthology/W15-3001 (cited on page 10).

Ondřej Bojar, Ondřej Dušek, Tom Kocmi, Jindřich Libovický, Michal Novák, Martin Popel, Roman Sudarikov, and Dušan Variš (2016). "CzEng 1.6: Enlarged Czech-English Parallel Corpus with Processing Tools Dockered". *Text, Speech, and Dialogue*. Edited by Petr Sojka, Aleš Horák, Ivan Kopeček, and Karel Pala. Cham: Springer International Publishing, pages 231–238. ISBN: 978-3-319-45510-5 (cited on page 42).

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi (2017). "Findings of the 2017 Conference on Machine Translation (WMT17)". *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*. Copenhagen, Denmark: Association for Computational Linguistics, pages 169–214. URL: http://www.aclweb.org/anthology/W17-4717 (cited on page 59).

Christian Buck and Philipp Koehn (2016). "Findings of the WMT 2016 Bilingual Document Alignment Shared Task". *Proceedings of the First Conference on Machine Translation*. Berlin, Germany: Association for Computational Linguistics,

pages 554–563. URL: http://www.aclweb.org/anthology/W/W16/W16-2347 (cited on page 91).

Chris Callison-Burch, Philipp Koehn, and Miles Osborne (2006). "Improved Statistical Machine Translation Using Paraphrases". *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*. New York City, USA: Association for Computational Linguistics, pages 17–24. URL: https://www.aclweb.org/anthology/N06-1003 (cited on page 51).

Marine Carpuat, Yogarshi Vyas, and Xing Niu (2017). "Detecting Cross-Lingual Semantic Divergence for Neural Machine Translation". *Proceedings of the First Workshop on Neural Machine Translation*. Vancouver: Association for Computational Linguistics, pages 69–79. DOI: 10.18653/v1/W17-3209. URL: https://www.aclweb.org/anthology/W17-3209 (cited on page 91).

Simon Carter, Wouter Weerkamp, and Manos Tsagkias (2013). "Microblog language identification: overcoming the limitations of short, unedited and idiomatic text". *Language Resources and Evaluation* 47.1, pages 195–215. ISSN: 1574020X, 15728412. URL: http://www.jstor.org/stable/42637260 (cited on page 78).

Isaac Caswell, Ciprian Chelba, and David Grangier (2019). "Tagged Back-Translation". *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*. Florence, Italy: Association for Computational Linguistics, pages 53–63. DOI: 10.18653/v1/W19-5206. URL: https://www.aclweb.org/anthology/W19-5206 (cited on page 33).

BIBLIOGRAPHY

Isaac Caswell, Theresa Breiner, Daan van Esch, and Ankur Bapna (2020). "Language ID in the Wild: Unexpected Challenges on the Path to a Thousand-Language Web Text Corpus". *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, pages 6588–6608. DOI: 10.18653/v1/2020.coling-main.579. URL: https://www.aclweb.org/anthology/2020.coling-main.579 (cited on page 78).

Isaac Caswell, Julia Kreutzer, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Javier Ortiz Suárez, Iroro Orife, Kelechi Ogueji, Rubungo Andre Niyongabo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi (2021). *Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets*. arXiv: 2103.12028 [cs.CL]. URL: https://arxiv.org/abs/2103.12028 (cited on pages 96, 100).

BIBLIOGRAPHY

Mauro Cettolo, Christian Girardi, and Marcello Federico (2012). "WIT^3: Web Inventory of Transcribed and Translated Talks". *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*. Trento, Italy, pages 261–268 (cited on page 60).

Sin-wai Chan (2002). *Translation and Information Technology*. Translation studies. Chinese University Press. ISBN: 9789629960773 (cited on page 2).

Vishrav Chaudhary, Yuqing Tang, Francisco Guzmán, Holger Schwenk, and Philipp Koehn (2019). "Low-Resource Corpus Filtering Using Multilingual Sentence Embeddings". *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*. Florence, Italy: Association for Computational Linguistics, pages 261–266. DOI: 10.18653/v1/W19-5435. URL: https://www.aclweb.org/anthology/W19-5435 (cited on page 94).

Colin Cherry and George Foster (2012). "Batch Tuning Strategies for Statistical Machine Translation". *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Montréal, Canada: Association for Computational Linguistics, pages 427–436. URL: https://www.aclweb.org/anthology/N12-1047 (cited on page 82).

David Chiang (2007). "Hierarchical Phrase-Based Translation". *Computational Linguistics* 33.2, pages 201–228. DOI: 10.1162/coli.2007.33.2.201. URL: https://www.aclweb.org/anthology/J07-2003 (cited on pages 13, 82).

BIBLIOGRAPHY

David Chiang, Kevin Knight, and Wei Wang (2009). "11,001 New Features for Statistical Machine Translation". *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Boulder, Colorado: Association for Computational Linguistics, pages 218–226. URL: https://www.aclweb.org/anthology/N09-1025 (cited on page 82).

Rohan Chitnis and John DeNero (2015). "Variable-Length Word Encodings for Neural Translation Models". *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, pages 2088–2093. DOI: 10.18653/v1/D15-1249. URL: https://www.aclweb.org/anthology/D15-1249 (cited on page 20).

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio (2014). "On the Properties of Neural Machine Translation: Encoder–Decoder Approaches". *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*. Doha, Qatar: Association for Computational Linguistics, pages 103–111. DOI: 10.3115/v1/W14-4012. URL: https://www.aclweb.org/anthology/W14-4012 (cited on page 15).

Julian Chow, Lucia Specia, and Pranava Madhyastha (2019). "WMDO: Fluency-based Word Mover's Distance for Machine Translation Evaluation". *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*. Florence, Italy: Association for Computational Linguistics, pages 494–500. DOI:

10.18653/v1/W19-5356. URL: https://www.aclweb.org/anthology/W19-5356 (cited on page 22).

Chenhui Chu, Raj Dabre, and Sadao Kurohashi (2017). "An Empirical Comparison of Domain Adaptation Methods for Neural Machine Translation". *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Vancouver, Canada: Association for Computational Linguistics, pages 385–391. DOI: 10.18653/v1/P17-2061. URL: http://www.aclweb.org/anthology/P17-2061 (cited on page 71).

Ryan Cotterell, Ekaterina Vylomova, Huda Khayrallah, Christo Kirov, and David Yarowsky (2017). "Paradigm Completion for Derivational Morphology". *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, pages 714–720. DOI: 10.18653/v1/D17-1074. URL: https://www.aclweb.org/anthology/D17-1074 (cited on page 5).

Josep Maria Crego, Jungi Kim, Guillaume Klein, Anabel Rebollo, Kathy Yang, Jean Senellart, Egor Akhanov, Patrice Brunelle, Aurélien Coquard, Yongchao Deng, Satoshi Enoue, Chiyo Geiss, Joshua Johanson, Ardas Khalsa, Raoum Khiari, Byeongil Ko, Catherine Kobus, Jean Lorieux, Leidiana Martins, Dang-Chuan Nguyen, Alexandra Priori, Thomas Riccardi, Natalia Segal, Christophe Servan, Cyril Tiquet, Bo Wang, Jin Yang, Dakun Zhang, Jing Zhou, and Peter Zoldan (2016). "SYSTRAN's Pure Neural Machine Translation Systems". *CoRR*

abs/1610.05540. arXiv: 1610.05540. URL: http://arxiv.org/abs/1610.05540 (cited on page 28).

Lei Cui, Dongdong Zhang, Shujie Liu, Mu Li, and Ming Zhou (2013). "Bilingual Data Cleaning for SMT using Graph-based Random Walk". *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Sofia, Bulgaria: Association for Computational Linguistics, pages 340–345. URL: https://www.aclweb.org/anthology/P13-2061 (cited on page 91).

Anna Currey, Antonio Valerio Miceli Barone, and Kenneth Heafield (2017). "Copied Monolingual Data Improves Low-Resource Neural Machine Translation". *Proceedings of the Second Conference on Machine Translation*. Copenhagen, Denmark: Association for Computational Linguistics, pages 148–156. DOI: 10.18653/v1/W17-4715. URL: https://www.aclweb.org/anthology/W17-4715 (cited on pages 33, 92).

Praveen Dakwale and Christof Monz (2017). "Fine-tuning for neural machine translation with limited degradation across in-and out-of-domain data". *Proceedings of the XVI Machine Translation Summit*, page 117 (cited on pages 27, 50, 72).

Jacob Devlin, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard Schwartz, and John Makhoul (2014). "Fast and Robust Neural Network Joint Models for Statistical Machine Translation". *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Baltimore, Maryland:

Association for Computational Linguistics, pages 1370–1380. DOI: 10.3115/v1/P14-1129. URL: https://www.aclweb.org/anthology/P14-1129 (cited on page 21).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pages 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://www.aclweb.org/anthology/N19-1423 (cited on page 32).

Shuoyang Ding, Kevin Duh, Huda Khayrallah, Philipp Koehn, and Matt Post (2016). "The JHU Machine Translation Systems for WMT 2016". *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*. Berlin, Germany: Association for Computational Linguistics, pages 272–280. DOI: 10.18653/v1/W16-2310. URL: https://www.aclweb.org/anthology/W16-2310 (cited on pages 5, 20, 81).

Shuoyang Ding, Huda Khayrallah, Philipp Koehn, Matt Post, Gaurav Kumar, and Kevin Duh (2017). "The JHU Machine Translation Systems for WMT 2017". *Proceedings of the Second Conference on Machine Translation*. Copenhagen, Denmark: Association for Computational Linguistics, pages 276–282. DOI: 10.18653/v1/W17-4724. URL: https://www.aclweb.org/anthology/W17-4724 (cited on pages 6, 81).

BIBLIOGRAPHY

Shuoyang Ding, Adithya Renduchintala, and Kevin Duh (2019). "A Call for Prudent Choice of Subword Merge Operations in Neural Machine Translation". *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*. Dublin, Ireland: European Association for Machine Translation, pages 204–213. URL: https://www.aclweb.org/anthology/W19-6620 (cited on page 20).

George Doddington (2002). "Automatic Evaluation of Machine Translation Quality Using N-Gram Co-Occurrence Statistics". *Proceedings of the Second International Conference on Human Language Technology Research*. HLT '02. San Diego, California: Morgan Kaufmann Publishers Inc., 138–145 (cited on page 22).

Tobias Domhan and Felix Hieber (2017). "Using Target-side Monolingual Data for Neural Machine Translation through Multi-task Learning". *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, pages 1500–1505. DOI: 10.18653/v1/D17-1158. URL: https://www.aclweb.org/anthology/D17-1158 (cited on page 52).

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang (2015). "Multi-Task Learning for Multiple Language Translation". *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Beijing, China: Association for Computational Linguistics, pages 1723–1732. DOI: 10.3115/

v1/P15-1166. URL: https://www.aclweb.org/anthology/P15-1166 (cited on page 30).

Markus Dreyer and Daniel Marcu (2012). "HyTER: Meaning-Equivalent Semantics for Translation Evaluation". *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Montréal, Canada: Association for Computational Linguistics, pages 162–171. URL: https://www.aclweb.org/anthology/N12-1017 (cited on page 37).

Kevin Duh (2018). *The Multitarget TED Talks Task*. http://www.cs.jhu.edu/~kevinduh/a/multitarget-tedtalks/ (cited on page 60).

Nadir Durrani, Alexander Fraser, Helmut Schmid, Hieu Hoang, and Philipp Koehn (2013). "Can Markov Models Over Minimal Translation Units Help Phrase-Based SMT?" *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Sofia, Bulgaria: Association for Computational Linguistics, pages 399–405. URL: https://www.aclweb.org/anthology/P13-2071 (cited on page 82).

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier (2018). "Understanding Back-Translation at Scale". *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pages 489–500. DOI: 10.18653/v1/D18-1045. URL: https://www.aclweb.org/anthology/D18-1045 (cited on pages 33, 100).

BIBLIOGRAPHY

Muhammad ElNokrashy, Amr Hendy, Mohamed Abdelghaffar, Mohamed Afify, Ahmed Tawfik, and Hany Hassan Awadalla (2020). "Score Combination for Improved Parallel Corpus Filtering for Low Resource Conditions". *Proceedings of the Fifth Conference on Machine Translation*. Online: Association for Computational Linguistics, pages 947–951. URL: https://www.aclweb.org/anthology/2020.wmt-1.106 (cited on page 95).

Miquel Esplà-Gomis, Víctor M. Sánchez-Cartagena, Jaume Zaragoza-Bernabeu, and Felipe Sánchez-Martínez (2020). "Bicleaner at WMT 2020: Universitat d'Alacant-Prompsit's submission to the parallel corpus filtering shared task". *Proceedings of the Fifth Conference on Machine Translation*. Online: Association for Computational Linguistics, pages 952–958. URL: https://www.aclweb.org/anthology/2020.wmt-1.107 (cited on page 95).

Marzieh Fadaee and Christof Monz (2018). "Back-Translation Sampling by Targeting Difficult Words in Neural Machine Translation". *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pages 436–446. DOI: 10.18653/v1/D18-1040. URL: https://www.aclweb.org/anthology/D18-1040 (cited on page 33).

Marzieh Fadaee, Arianna Bisazza, and Christof Monz (2017). "Data Augmentation for Low-Resource Neural Machine Translation". *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Vancouver, Canada: Association for Computational Linguistics, pages 567–573.

DOI: 10.18653/v1/P17-2090. URL: https://www.aclweb.org/anthology/P17-2090 (cited on pages 33, 51).

Angela Fan, Mike Lewis, and Yann Dauphin (2018). "Hierarchical Neural Story Generation". *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, pages 889–898. DOI: 10.18653/v1/P18-1082. URL: https://www.aclweb.org/anthology/P18-1082 (cited on page 43).

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio (2016). "Multi-Way, Multilingual Neural Machine Translation with a Shared Attention Mechanism". *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, pages 866–875. DOI: 10.18653/v1/N16-1101. URL: https://www.aclweb.org/anthology/N16-1101 (cited on page 30).

George Foster, Cyril Goutte, and Roland Kuhn (2010). "Discriminative Instance Weighting for Domain Adaptation in Statistical Machine Translation". *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Cambridge, MA: Association for Computational Linguistics, pages 451–459. URL: https://www.aclweb.org/anthology/D10-1044 (cited on page 27).

Markus Freitag and Yaser Al-Onaizan (2016). "Fast Domain Adaptation for Neural Machine Translation". *CoRR* abs/1612.06897 (cited on page 71).

BIBLIOGRAPHY

Philip Gage (1994). "A New Algorithm for Data Compression". *C Users J.* 12.2, 23–38. ISSN: 0898-9788 (cited on page 19).

Michel Galley and Christopher D. Manning (2008). "A Simple and Effective Hierarchical Phrase Reordering Model". *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Honolulu, Hawaii: Association for Computational Linguistics, pages 848–856. URL: https://www.aclweb.org/anthology/D08-1089 (cited on pages 13, 82).

Michel Galley, Mark Hopkins, Kevin Knight, and Daniel Marcu (2004). "What's in a translation rule?" *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*. Boston, Massachusetts, USA: Association for Computational Linguistics, pages 273–280. URL: https://www.aclweb.org/anthology/N04-1035 (cited on pages 13, 82).

Michel Galley, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang, and Ignacio Thayer (2006). "Scalable Inference and Training of Context-Rich Syntactic Translation Models". *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*. Sydney, Australia: Association for Computational Linguistics, pages 961–968. DOI: 10.3115/1220175.1220296. URL: https://www.aclweb.org/anthology/P06-1121 (cited on page 82).

BIBLIOGRAPHY

Jonas Gehring, Michael Auli, David Grangier, and Yann Dauphin (2017). "A Convolutional Encoder Model for Neural Machine Translation". *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, pages 123–135. DOI: 10.18653/v1/P17-1012. URL: https://www.aclweb.org/anthology/P17-1012 (cited on page 96).

Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio (2013). "An empirical investigation of catastrophic forgetting in gradient-based neural networks". *arXiv preprint arXiv:1312.6211* (cited on pages 55, 73).

Shuhao Gu and Yang Feng (2020). "Investigating Catastrophic Forgetting During Continual Training for Neural Machine Translation". *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, pages 4315–4326. DOI: 10.18653/v1/2020.coling-main.381. URL: https://www.aclweb.org/anthology/2020.coling-main.381 (cited on page 29).

Biman Gujral, Huda Khayrallah, and Philipp Koehn (2016). "Translation of Unknown Words in Low Resource Languages". *Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA)* (cited on page 5).

Çaglar Gülçehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loïc Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio (2015). "On Using

Monolingual Corpora in Neural Machine Translation". *CoRR* abs/1503.03535. arXiv: 1503.03535. URL: http://arxiv.org/abs/1503.03535 (cited on page 52).

Çaglar Gülçehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, and Yoshua Bengio (2017). "On Integrating a Language Model into Neural Machine Translation". *Comput. Speech Lang.* 45.C, pages 137–148. ISSN: 0885-2308. DOI: 10.1016/j.csl.2017.01.014. URL: https://doi.org/10.1016/j.csl.2017.01.014 (cited on page 52).

Rohit Gupta, Constantin Orăsan, and Josef van Genabith (2015). "ReVal: A Simple and Effective Machine Translation Evaluation Metric Based on Recurrent Neural Networks". *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, pages 1066–1072. DOI: 10.18653/v1/D15-1124. URL: https://www.aclweb.org/anthology/D15-1124 (cited on page 22).

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato (2019). "The FLORES Evaluation Datasets for Low-Resource Machine Translation: Nepali–English and Sinhala–English". *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pages 6098–6111. DOI: 10.18653/v1/D19-1632. URL: https://www.aclweb.org/anthology/D19-1632 (cited on pages 20, 42).

BIBLIOGRAPHY

Barry Haddow, Matthias Huck, Alexandra Birch, Nikolay Bogoychev, and Philipp Koehn (2015). "The Edinburgh/JHU Phrase-based Machine Translation Systems for WMT 2015". *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Lisbon, Portugal: Association for Computational Linguistics, pages 126–133. DOI: 10.18653/v1/W15-3013. URL: https://www.aclweb.org/anthology/W15-3013 (cited on page 81).

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou (2018). "Achieving Human Parity on Automatic Chinese to English News Translation". *CoRR* abs/1803.05567. arXiv: 1803.05567. URL: http://arxiv.org/abs/1803.05567 (cited on page 3).

Kenneth Heafield (2011). "KenLM: Faster and Smaller Language Model Queries". *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Edinburgh, Scotland: Association for Computational Linguistics, pages 187–197. URL: https://www.aclweb.org/anthology/W11-2123 (cited on page 82).

Kenneth Heafield, Hiroaki Hayashi, Yusuke Oda, Ioannis Konstas, Andrew Finch, Graham Neubig, Xian Li, and Alexandra Birch (2020). "Findings of the Fourth Workshop on Neural Generation and Translation". *Proceedings of the Fourth Workshop on Neural Generation and Translation*. Online: Association for

Computational Linguistics, pages 1–9. DOI: 10.18653/v1/2020.ngt-1.1. URL: https://www.aclweb.org/anthology/2020.ngt-1.1 (cited on page 2).

G. E. Hinton and R. R. Salakhutdinov (2006). "Reducing the Dimensionality of Data with Neural Networks". *Science* 313.5786, pages 504–507. ISSN: 0036-8075. DOI: 10.1126/science.1127647. eprint: http://science.sciencemag.org/content/313/5786/504.full.pdf. URL: http://science.sciencemag.org/content/313/5786/504 (cited on pages 25, 55).

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean (2015). *Distilling the Knowledge in a Neural Network*. arXiv: 1503.02531 [stat.ML] (cited on pages 17, 50, 69).

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn (2018). "Iterative Back-Translation for Neural Machine Translation". *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*. Melbourne, Australia: Association for Computational Linguistics, pages 18–24. DOI: 10.18653/v1/W18-2703. URL: https://www.aclweb.org/anthology/W18-2703 (cited on page 33).

J. Edward Hu, Huda Khayrallah, Ryan Culkin, Patrick Xia, Tongfei Chen, Matt Post, and Benjamin Van Durme (2019a). "Improved Lexically Constrained Decoding for Translation and Monolingual Rewriting". *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pages 839–850. DOI: 10.18653/v1/N19-

1090. URL: https://www.aclweb.org/anthology/N19-1090 (cited on pages 5, 50, 51, 53).

J. Edward Hu, Abhinav Singh, Nils Holzenberger, Matt Post, and Benjamin Van Durme (2019b). "Large-Scale, Diverse, Paraphrastic Bitexts via Sampling and Clustering". *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*. Hong Kong, China: Association for Computational Linguistics, pages 44–54. DOI: 10.18653/v1/K19-1005. URL: https://www.aclweb.org/anthology/K19-1005 (cited on pages 42, 53).

J. Edward Hu, Rachel Rudinger, Matt Post, and Benjamin Van Durme (2019c). "ParaBank: Monolingual Bitext Generation and Sentential Paraphrasing via Lexically-constrained Neural Machine Translation". *Proceedings of AAAI*. DOI: 10.1609/aaai.v33i01.33016521. URL: https://aaai.org/ojs/index.php/AAAI/article/view/4618 (cited on page 53).

Liang Huang and David Chiang (2007). "Forest Rescoring: Faster Decoding with Integrated Language Models". *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Prague, Czech Republic: Association for Computational Linguistics, pages 144–151. URL: https://www.aclweb.org/anthology/P07-1019 (cited on page 82).

David A. Huffman (1952). "A Method for the Construction of Minimum-Redundancy Codes". *Proceedings of the IRE* 40.9, pages 1098–1101. DOI: 10.1109/JRPROC.1952.273898 (cited on page 20).

BIBLIOGRAPHY

John Hutchins (1997). "From First Conception to First Demonstration: the Nascent Years of Machine Translation, 1947-1954. A Chronology". *Machine Translation* 12.3. Full publication date: 1997, pages 195–252. URL: http://www.jstor.org/stable/40027329 (cited on page 11).

Kenji Imamura and Eiichiro Sumita (2016). "Multi-Domain Adaptation for Statistical Machine Translation Based on Feature Augmentation". *Proceedings of the 12th Conference of the Association for Machine Translation in the Americas,* Austin, Texas, USA (cited on page 27).

Sébastien Jean, Orhan Firat, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio (2015a). "Montreal Neural Machine Translation Systems for WMT'15". *Proceedings of the Tenth Workshop on Statistical Machine Translation.* Lisbon, Portugal: Association for Computational Linguistics, pages 134–140. DOI: 10.18653/v1/W15-3014. URL: https://www.aclweb.org/anthology/W15-3014 (cited on page 10).

Sébastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio (2015b). "On Using Very Large Target Vocabulary for Neural Machine Translation". *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers).* Beijing, China: Association for Computational Linguistics, pages 1–10. DOI: 10.3115/v1/P15-1001. URL: https://www.aclweb.org/anthology/P15-1001 (cited on page 18).

BIBLIOGRAPHY

Jing Jiang and ChengXiang Zhai (2007). "Instance Weighting for Domain Adaptation in NLP". *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Prague, Czech Republic: Association for Computational Linguistics, pages 264–271. URL: https://www.aclweb.org/anthology/P07-1034 (cited on page 27).

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean (2017). "Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation". *Transactions of the Association for Computational Linguistics* 5, pages 339–351. DOI: 10.1162/tacl_a_00065. URL: https://www.aclweb.org/anthology/Q17-1024 (cited on page 31).

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy (2020). "SpanBERT: Improving Pre-training by Representing and Predicting Spans". *Transactions of the Association for Computational Linguistics* 8, pages 64–77. DOI: 10.1162/tacl_a_00300. URL: https://www.aclweb.org/anthology/2020.tacl-1.5 (cited on page 32).

Marcin Junczys-Dowmunt (2012). "A Phrase Table without Phrases: Rank Encoding for Better Phrase Table Compression". *Proceedings of the 16th Annual conference of the European Association for Machine Translation*. Trento, Italy: European Association for Machine Translation, pages 245–252. URL: https://www.aclweb.org/anthology/2012.eamt-1.58 (cited on page 82).

BIBLIOGRAPHY

Marcin Junczys-Dowmunt (2018). "Dual Conditional Cross-Entropy Filtering of Noisy Parallel Corpora". *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*. Belgium, Brussels: Association for Computational Linguistics, pages 888–895. DOI: 10.18653/v1/W18-6478. URL: https://www.aclweb.org/anthology/W18-6478 (cited on page 94).

Marcin Junczys-Dowmunt (2019). "Microsoft Translator at WMT 2019: Towards Large-Scale Document-Level Neural Machine Translation". *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*. Florence, Italy: Association for Computational Linguistics, pages 225–233. DOI: 10.18653/v1/W19-5321. URL: https://www.aclweb.org/anthology/W19-5321 (cited on page 15).

Marcin Junczys-Dowmunt, Tomasz Dwojak, and Rico Sennrich (2016). "The AMU-UEDIN Submission to the WMT16 News Translation Task: Attention-based NMT Models as Feature Functions in Phrase-based SMT". *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*. Berlin, Germany: Association for Computational Linguistics, pages 319–325. DOI: 10.18653/v1/W16-2316. URL: https://www.aclweb.org/anthology/W16-2316 (cited on page 21).

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch (2018). "Marian:

Fast Neural Machine Translation in C++". *arXiv preprint arXiv:1804.00344*. URL: https://arxiv.org/abs/1804.00344 (cited on page 81).

Nal Kalchbrenner and Phil Blunsom (2013). "Recurrent Continuous Translation Models". *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington, USA: Association for Computational Linguistics, pages 1700–1709. URL: https://www.aclweb.org/anthology/D13-1176 (cited on pages 10, 15, 96).

Huda Khayrallah and Philipp Koehn (2018). "On the Impact of Various Types of Noise on Neural Machine Translation". *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*. Melbourne, Australia: Association for Computational Linguistics, pages 74–83. DOI: 10.18653/v1/W18-2709. URL: https://www.aclweb.org/anthology/W18-2709 (cited on pages 7, 22, 75).

Huda Khayrallah and João Sedoc (2020). "SMRT Chatbots: Improving Non-Task-Oriented Dialog with Simulated Multiple Reference Training". *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, pages 4489–4505. DOI: 10.18653/v1/2020.findings-emnlp.403. URL: https://www.aclweb.org/anthology/2020.findings-emnlp.403 (cited on pages 8, 37).

Huda Khayrallah and João Sedoc (2021). "Measuring the 'I don't know' Problem through the Lens of Gricean Quantity". *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics:*

Human Language Technologies. Online: Association for Computational Linguistics, pages 5659–5670. URL: https://www.aclweb.org/anthology/2021.naacl-main.450 (cited on page 8).

Huda Khayrallah, Gaurav Kumar, Kevin Duh, Matt Post, and Philipp Koehn (2017). "Neural Lattice Search for Domain Adaptation in Machine Translation". *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Taipei, Taiwan: Asian Federation of Natural Language Processing, pages 20–25. URL: https://www.aclweb.org/anthology/I17-2004 (cited on pages 6, 21, 28).

Huda Khayrallah, Brian Thompson, Kevin Duh, and Philipp Koehn (2018a). "Regularized Training Objective for Continued Training for Domain Adaptation in Neural Machine Translation". *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*. Melbourne, Australia: Association for Computational Linguistics, pages 36–44. DOI: 10.18653/v1/W18-2705. URL: https://www.aclweb.org/anthology/W18-2705 (cited on pages 6, 50, 55).

Huda Khayrallah, Hainan Xu, and Philipp Koehn (2018b). "The JHU Parallel Corpus Filtering Systems for WMT 2018". *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*. Belgium, Brussels: Association for Computational Linguistics, pages 896–899. DOI: 10.18653/v1/W18-6479. URL: https://www.aclweb.org/anthology/W18-6479 (cited on page 7).

BIBLIOGRAPHY

Huda Khayrallah, Rebecca Knowles, Kevin Duh, and Matt Post (2019). "An Interactive Teaching Tool for Introducing Novices to Machine Translation". *Proceedings of the 50th ACM Technical Symposium on Computer Science Education*. SIGCSE '19. Minneapolis, MN, USA: Association for Computing Machinery, page 1276. ISBN: 9781450358903. DOI: 10.1145/3287324.3293840. URL: https://doi.org/10.1145/3287324.3293840 (cited on page 7).

Huda Khayrallah, Brian Thompson, Matt Post, and Philipp Koehn (2020a). "Simulated multiple reference training improves low-resource machine translation". *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pages 82–89. DOI: 10.18653/v1/2020.emnlp-main.7. URL: https://www.aclweb.org/anthology/2020.emnlp-main.7 (cited on pages 5, 8, 37, 100).

Huda Khayrallah, Jacob Bremerman, Arya D. McCarthy, Kenton Murray, Winston Wu, and Matt Post (2020b). "The JHU Submission to the 2020 Duolingo Shared Task on Simultaneous Translation and Paraphrase for Language Education". *Proceedings of the Fourth Workshop on Neural Generation and Translation*. Online: Association for Computational Linguistics, pages 188–197. DOI: 10.18653/v1/2020.ngt-1.22. URL: https://www.aclweb.org/anthology/2020.ngt-1.22 (cited on page 8).

Yoon Kim and Alexander M. Rush (2016). "Sequence-Level Knowledge Distillation". *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics,

pages 1317–1327. DOI: 10.18653/v1/D16-1139. URL: https://www.aclweb.org/anthology/D16-1139 (cited on pages 17, 50, 69).

Diederik P. Kingma and Jimmy Ba (2015). "Adam: A Method for Stochastic Optimization". *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Edited by Yoshua Bengio and Yann LeCun. URL: http://arxiv.org/abs/1412.6980 (cited on page 42).

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell (2017). "Overcoming catastrophic forgetting in neural networks". *Proceedings of the National Academy of Sciences* 114.13, pages 3521–3526. ISSN: 0027-8424. DOI: 10.1073/pnas.1611835114. eprint: http://www.pnas.org/content/114/13/3521.full.pdf. URL: http://www.pnas.org/content/114/13/3521 (cited on page 73).

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush (2017). "OpenNMT: Open-Source Toolkit for Neural Machine Translation". *Proc. ACL*. DOI: 10.18653/v1/P17-4012. URL: https://doi.org/10.18653/v1/P17-4012 (cited on page 61).

BIBLIOGRAPHY

Rebecca Knowles and Philipp Koehn (2016). "Neural Interactive Translation Prediction". *Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA)* (cited on page 2).

Catherine Kobus, Josep Crego, and Jean Senellart (2017). "Domain Control for Neural Machine Translation". *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*. Varna, Bulgaria: INCOMA Ltd., pages 372–378. DOI: 10.26615/978-954-452-049-6_049. URL: https://doi.org/10.26615/978-954-452-049-6_049 (cited on page 28).

Philipp Koehn (2004). "Statistical Significance Tests for Machine Translation Evaluation". *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. Barcelona, Spain: Association for Computational Linguistics, pages 388–395. URL: https://www.aclweb.org/anthology/W04-3250 (cited on page 44).

Philipp Koehn (2005). "Europarl: A Parallel Corpus for Statistical Machine Translation". *Proceedings of the Tenth Machine Translation Summit (MT Summit X)*. Phuket, Thailand. URL: http://mt-archive.info/MTS-2005-Koehn.pdf (cited on pages 34, 59, 79, 82).

Philipp Koehn (2009). *Statistical Machine Translation*. Cambridge University Press. DOI: 10.1017/CBO9780511815829 (cited on pages 13, 24).

Philipp Koehn and Kevin Knight (2003). "Empirical Methods for Compound Splitting". *10th Conference of the European Chapter of the Association for Computational*

Linguistics. Budapest, Hungary: Association for Computational Linguistics. URL: https://www.aclweb.org/anthology/E03-1076 (cited on page 18).

Philipp Koehn and Rebecca Knowles (2017). "Six Challenges for Neural Machine Translation". *Proceedings of the First Workshop on Neural Machine Translation*. Vancouver: Association for Computational Linguistics, pages 28–39. DOI: 10.18653/v1/W17-3204. URL: https://www.aclweb.org/anthology/W17-3204 (cited on pages 4, 22, 60, 98).

Philipp Koehn, Franz J. Och, and Daniel Marcu (2003). "Statistical Phrase-Based Translation". *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 127–133. URL: https://www.aclweb.org/anthology/N03-1017 (cited on page 13).

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst (2007). "Moses: Open Source Toolkit for Statistical Machine Translation". *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*. Prague, Czech Republic: Association for Computational Linguistics, pages 177–180. URL: https://www.aclweb.org/anthology/P07-2045 (cited on page 81).

BIBLIOGRAPHY

Philipp Koehn, Huda Khayrallah, Kenneth Heafield, and Mikel L. Forcada (2018). "Findings of the WMT 2018 Shared Task on Parallel Corpus Filtering". *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*. Belgium, Brussels: Association for Computational Linguistics, pages 726–739. DOI: 10.18653/v1/W18-6453. URL: https://www.aclweb.org/anthology/W18-6453 (cited on pages 7, 35, 93).

Philipp Koehn, Francisco Guzmán, Vishrav Chaudhary, and Juan Pino (2019). "Findings of the WMT 2019 Shared Task on Parallel Corpus Filtering for Low-Resource Conditions". *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*. Florence, Italy: Association for Computational Linguistics, pages 54–72. DOI: 10.18653/v1/W19-5404. URL: https://www.aclweb.org/anthology/W19-5404 (cited on page 94).

Philipp Koehn, Vishrav Chaudhary, Ahmed El-Kishky, Naman Goyal, Peng-Jen Chen, and Francisco Guzmán (2020). "Findings of the WMT 2020 Shared Task on Parallel Corpus Filtering and Alignment". *Proceedings of the Fifth Conference on Machine Translation*. Online: Association for Computational Linguistics, pages 726–742. URL: https://www.aclweb.org/anthology/2020.wmt-1.78 (cited on pages 94, 95).

Taku Kudo (2018). "Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates". *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, pages 66–75.

DOI: 10.18653/v1/P18-1007. URL: https://www.aclweb.org/anthology/P18-1007 (cited on pages 19, 20).

Taku Kudo and John Richardson (2018). "SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing". *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Brussels, Belgium: Association for Computational Linguistics, pages 66–71. DOI: 10.18653/v1/D18-2012. URL: https://www.aclweb.org/anthology/D18-2012 (cited on pages 19, 43).

Gaurav Kumar, George Foster, Colin Cherry, and Maxim Krikun (2019). "Reinforcement Learning based Curriculum Optimization for Neural Machine Translation". *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pages 2054–2061. DOI: 10.18653/v1/N19-1208. URL: https://www.aclweb.org/anthology/N19-1208 (cited on page 101).

Gaurav Kumar, Philipp Koehn, and Sanjeev Khudanpur (2021a). *Learning Feature Weights using Reward Modeling for Denoising Parallel Corpora*. arXiv: 2103.06968 [cs.CL] (cited on page 101).

Gaurav Kumar, Philipp Koehn, and Sanjeev Khudanpur (2021b). *Learning Policies for Multilingual Training of Neural Machine Translation Systems*. arXiv: 2103.06964 [cs.CL] (cited on page 101).

BIBLIOGRAPHY

Shankar Kumar and William Byrne (2004). "Minimum Bayes-Risk Decoding for Statistical Machine Translation". *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*. Boston, Massachusetts, USA: Association for Computational Linguistics, pages 169–176. URL: https://www.aclweb.org/anthology/N04-1022 (cited on page 82).

Alon Lavie and Abhaya Agarwal (2007). "METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments". *Proceedings of the Second Workshop on Statistical Machine Translation*. Prague, Czech Republic: Association for Computational Linguistics, pages 228–231. URL: https://www.aclweb.org/anthology/W07-0734 (cited on page 22).

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer (2020). "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension". *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pages 7871–7880. DOI: 10.18653/v1/2020.acl-main.703. URL: https://www.aclweb.org/anthology/2020.acl-main.703 (cited on page 31).

William Lewis (2010). "Haitian Creole: How to Build and Ship an MT Engine from Scratch in 4 days, 17 hours, & 30 minutes". *Proceedings of the 14th Annual conference of the European Association for Machine Translation*. Saint Raphaël,

France: European Association for Machine Translation. URL: https://www.aclweb.org/anthology/2010.eamt-1.37 (cited on page 101).

Zichao Li, Xin Jiang, Lifeng Shang, and Hang Li (2018). "Paraphrase Generation with Deep Reinforcement Learning". *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pages 3865–3878. DOI: 10.18653/v1/D18-1421. URL: https://www.aclweb.org/anthology/D18-1421 (cited on page 53).

Pierre Lison and Jörg Tiedemann (2016). "OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles". *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia: European Language Resources Association (ELRA), pages 923–929. URL: https://www.aclweb.org/anthology/L16-1147 (cited on page 84).

Zhengyuan Liu, Ke Shi, and Nancy Chen (2020). "Multilingual Neural RST Discourse Parsing". *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, pages 6730–6738. DOI: 10.18653/v1/2020.coling-main.591. URL: https://www.aclweb.org/anthology/2020.coling-main.591 (cited on pages 31, 32).

Chi-kiu Lo (2017). "MEANT 2.0: Accurate semantic MT evaluation for any output language". *Proceedings of the Second Conference on Machine Translation*.

Copenhagen, Denmark: Association for Computational Linguistics, pages 589–597. DOI: 10.18653/v1/W17-4767. URL: https://www.aclweb.org/anthology/W17-4767 (cited on page 22).

Chi-kiu Lo (2019). "YiSi - a Unified Semantic MT Quality Evaluation and Estimation Metric for Languages with Different Levels of Available Resources". *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*. Florence, Italy: Association for Computational Linguistics, pages 507–513. DOI: 10.18653/v1/W19-5358. URL: https://www.aclweb.org/anthology/W19-5358 (cited on page 22).

Chi-kiu Lo and Dekai Wu (2011). "MEANT: An inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility based on semantic roles". *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, pages 220–229. URL: https://www.aclweb.org/anthology/P11-1023 (cited on page 22).

Minh-Thang Luong and Christopher D. Manning (2015). "Stanford Neural Machine Translation Systems for Spoken Language Domain". *International Workshop on Spoken Language Translation*. Da Nang, Vietnam (cited on pages 25, 55, 70).

Thang Luong, Ilya Sutskever, Quoc Le, Oriol Vinyals, and Wojciech Zaremba (2015). "Addressing the Rare Word Problem in Neural Machine Translation". *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the*

7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Beijing, China: Association for Computational Linguistics, pages 11–19. DOI: 10.3115/v1/P15-1002. URL: https://www.aclweb.org/anthology/P15-1002 (cited on page 18).

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham (2019). "Results of the WMT19 Metrics Shared Task: Segment-Level and Strong MT Systems Pose Big Challenges". *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*. Florence, Italy: Association for Computational Linguistics, pages 62–90. DOI: 10.18653/v1/W19-5302. URL: https://www.aclweb.org/anthology/W19-5302 (cited on page 23).

Nitin Madnani, Necip Fazil Ayan, Philip Resnik, and Bonnie Dorr (2007). "Using Paraphrases for Parameter Tuning in Statistical Machine Translation". *Proceedings of the Second Workshop on Statistical Machine Translation*. Prague, Czech Republic: Association for Computational Linguistics, pages 120–127. URL: https://www.aclweb.org/anthology/W07-0716 (cited on page 51).

Nitin Madnani, Philip Resnik, Bonnie J. Dorr, and Richard Schwartz (2008). "Are Multiple Reference Translations Necessary? Investigating the Value of Paraphrased Reference Translations in Parameter Optimization". *Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas*. Waikiki, Hawaii (cited on page 51).

BIBLIOGRAPHY

Marianna Martindale and Marine Carpuat (2018). "Fluency Over Adequacy: A Pilot Study in Measuring User Trust in Imperfect MT". *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*. Boston, MA: Association for Machine Translation in the Americas, pages 13–25. URL: https://www.aclweb.org/anthology/W18-1803 (cited on page 21).

Marianna Martindale, Marine Carpuat, Kevin Duh, and Paul McNamee (2019). "Identifying Fluently Inadequate Output in Neural and Statistical Machine Translation". *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*. Dublin, Ireland: European Association for Machine Translation, pages 233–243. URL: https://www.aclweb.org/anthology/W19-6623 (cited on page 21).

Marianna J. Martindale (2020). "Responsible 'Gist' MT Use in the Age of Neural MT". *Workshop on the Impact of Machine Translation (iMpacT 2020)*. Virtual: Association for Machine Translation in the Americas, pages 18–45. URL: https://www.aclweb.org/anthology/2020.amta-impact.2 (cited on page 21).

Yuval Marton, Chris Callison-Burch, and Philip Resnik (2009). "Improved Statistical Machine Translation Using Monolingually-Derived Paraphrases". *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, pages 381–390. URL: https://www.aclweb.org/anthology/D09-1040 (cited on page 51).

BIBLIOGRAPHY

Nitika Mathur, Timothy Baldwin, and Trevor Cohn (2019). "Putting Evaluation in Context: Contextual Embeddings Improve Machine Translation Evaluation". *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pages 2799–2808. DOI: 10.18653/v1/P19-1269. URL: https://www.aclweb.org/anthology/P19-1269 (cited on page 22).

Nitika Mathur, Timothy Baldwin, and Trevor Cohn (2020). "Tangled up in BLEU: Reevaluating the Evaluation of Automatic Machine Translation Evaluation Metrics". *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pages 4984–4997. DOI: 10.18653/v1/2020.acl-main.448. URL: https://www.aclweb.org/anthology/2020.acl-main.448 (cited on page 23).

Spyros Matsoukas, Antti-Veikko I. Rosti, and Bing Zhang (2009). "Discriminative Corpus Weight Estimation for Machine Translation". *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, pages 708–717. URL: https://www.aclweb.org/anthology/D09-1074 (cited on page 27).

Haitao Mi, Zhiguo Wang, and Abe Ittycheriah (2016). "Vocabulary Manipulation for Neural Machine Translation". *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Berlin, Germany: Association for Computational Linguistics, pages 124–129. DOI: 10.

18653/v1/P16-2021. URL: https://www.aclweb.org/anthology/P16-2021 (cited on page 21).

Antonio Valerio Miceli Barone, Barry Haddow, Ulrich Germann, and Rico Sennrich (2017). "Regularization techniques for fine-tuning in neural machine translation". *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, pages 1489–1494. DOI: 10.18653/v1/D17-1156. URL: https://www.aclweb.org/anthology/D17-1156 (cited on page 70).

Antonio Valerio Miceli Barone, Barry Haddow, Ulrich Germann, and Rico Sennrich (2017). "Regularization techniques for fine-tuning in neural machine translation". *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, pages 1489–1494. URL: https://www.aclweb.org/anthology/D17-1156 (cited on page 71).

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton (2019). "When does label smoothing help?" *Advances in Neural Information Processing Systems*. Edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Volume 32. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2019/file/f1748d6b0fd9d439f71450117eba2725-Paper.pdf (cited on page 18).

Toan Q. Nguyen and David Chiang (2017). "Transfer Learning across Low-Resource, Related Languages for Neural Machine Translation". *Proceedings of the Eighth*

International Joint Conference on Natural Language Processing (Volume 2: Short Papers). Taipei, Taiwan: Asian Federation of Natural Language Processing, pages 296–301. URL: https://www.aclweb.org/anthology/I17-2050 (cited on page 30).

Sonja Nießen and Hermann Ney (2000). "Improving SMT quality with morpho-syntactic analysis". *COLING 2000 Volume 2: The 18th International Conference on Computational Linguistics*. URL: https://www.aclweb.org/anthology/C00-2162 (cited on page 18).

Mohammad Norouzi, Samy Bengio, zhifeng Chen, Navdeep Jaitly, Mike Schuster, Yonghui Wu, and Dale Schuurmans (2016). "Reward Augmented Maximum Likelihood for Neural Structured Prediction". *Advances in Neural Information Processing Systems 29*. Edited by D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett. Curran Associates, Inc., pages 1723–1731. URL: http://papers.nips.cc/paper/6547-reward-augmented-maximum-likelihood-for-neural-structured-prediction.pdf (cited on pages 33, 52).

Miles Osborne (2010). "Statistical Machine Translation". *Encyclopedia of Machine Learning*. Edited by Claude Sammut and Geoffrey I. Webb. Boston, MA: Springer US, pages 912–915. ISBN: 978-0-387-30164-8. DOI: 10.1007/978-0-387-30164-8_783. URL: https://doi.org/10.1007/978-0-387-30164-8_783 (cited on page 13).

Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato (2018). "Analyzing Uncertain hine Translation". *CoRR* abs/1803.00047. arXiv: 1803.00047. URL: http://arxiv.org/abs/1803.00047 (cited on page 93).

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli (2019). "fairseq: A Fast, Extensible Toolkit for Sequence Modeling". *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*. Minneapolis, Minnesota: Association for Computational Linguistics, pages 48–53. DOI: 10.18653/v1/N19-4009. URL: https://www.aclweb.org/anthology/N19-4009 (cited on page 42).

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu (2002). "Bleu: a Method for Automatic Evaluation of Machine Translation". *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, pages 311–318. DOI: 10.3115/1073083.1073135. URL: https://www.aclweb.org/anthology/P02-1040 (cited on pages 22, 85).

Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton (2017). "Regularizing neural networks by penalizing confident output distributions". *International Conference on Learning Representations (ICLR) - Workshop Track*. URL: https://openreview.net/group?id=ICLR.cc/2017/workshop (cited on pages 52, 70).

BIBLIOGRAPHY

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer (2018). "Deep Contextualized Word Representations". *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pages 2227–2237. DOI: 10.18653/v1/N18-1202. URL: https://www.aclweb.org/anthology/N18-1202 (cited on page 31).

Maja Popović (2015). "chrF: character n-gram F-score for automatic MT evaluation". *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Lisbon, Portugal: Association for Computational Linguistics, pages 392–395. DOI: 10.18653/v1/W15-3049. URL: https://www.aclweb.org/anthology/W15-3049 (cited on page 22).

Maja Popović (2017). "chrF++: words helping character n-grams". *Proceedings of the Second Conference on Machine Translation*. Copenhagen, Denmark: Association for Computational Linguistics, pages 612–618. DOI: 10.18653/v1/W17-4770. URL: https://www.aclweb.org/anthology/W17-4770 (cited on page 22).

Matt Post (2018). "A Call for Clarity in Reporting BLEU Scores". *Proceedings of the Third Conference on Machine Translation: Research Papers*. Brussels, Belgium: Association for Computational Linguistics, pages 186–191. DOI: 10.18653/v1/W18-6319. URL: https://www.aclweb.org/anthology/W18-6319 (cited on pages 23, 44).

BIBLIOGRAPHY

Alexandre Rafalovitch and Robert Dale (2009). "United Nations General Assembly Resolutions: A Six-Language Parallel Corpus". *Proceedings of the Twelfth Machine Translation Summit (MT Summit XII)*. Ottawa, Ontario, Canada: International Association for Machine Translation. URL: http://www.uncorpora.org/Rafalovitch_Dale_MT_Summit_2009.pdf (cited on page 34).

Spencer Rarrick, Chris Quirk, and Will Lewis (2011). "MT Detection in Web-Scraped Parallel Corpora". *Proceedings of the 13th Machine Translation Summit (MT Summit XIII)*. Xiamen, China: International Association for Machine Translation, pages 422–430. URL: http://www.mt-archive.info/MTS-2011-Rarrick.pdf (cited on page 91).

Philip Resnik (1999). "Mining the Web for Bilingual Text". *Proceedings of the 37th Annual Meeting of the Association of Computational Linguistics (ACL)*. URL: http://acl.ldc.upenn.edu/P/P99/P99-1068.pdf (cited on page 34).

Anthony Rousseau, Fethi Bougares, Paul Deléglise, Holger Schwenk, and Yannick Esteve (2011). "LIUMS Systems for the IWSLT 2011 Speech Translation Tasks". San Francisco, CA. URL: https://www.isca-speech.org/archive/iwslt_11/papers/sltb_079.pdf (cited on page 27).

Carl Rubino (2018). "Keynote: Setting up a Machine Translation Program for IARPA". *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 2: User Track)*. Boston, MA: Association for Machine

Translation in the Americas. URL: https://www.aclweb.org/anthology/W18-1902 (cited on page 43).

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams (1986). "Learning representations by back-propagating errors". *Nature* 323.6088, pages 533–536. ISSN: 1476-4687. DOI: 10.1038/323533a0. URL: https://doi.org/10.1038/323533a0 (cited on page 15).

Danielle Saunders (2021). "Domain Adaptation and Multi-Domain Adaptation for Neural Machine Translation: A Survey". *CoRR* abs/2104.06951. arXiv: 2104.06951. URL: https://arxiv.org/abs/2104.06951 (cited on page 28).

Holger Schwenk (2007). "Continuous Space Language Models". *Comput. Speech Lang.* 21.3, pages 492–518. ISSN: 0885-2308. DOI: 10.1016/j.csl.2006.09.003. URL: http://dx.doi.org/10.1016/j.csl.2006.09.003 (cited on page 52).

Holger Schwenk (2012). "Continuous Space Translation Models for Phrase-Based Statistical Machine Translation". *Proceedings of COLING 2012: Posters*. Mumbai, India: The COLING 2012 Organizing Committee, pages 1071–1080. URL: https://www.aclweb.org/anthology/C12-2104 (cited on page 52).

Thibault Sellam, Dipanjan Das, and Ankur Parikh (2020). "BLEURT: Learning Robust Metrics for Text Generation". *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pages 7881–7892. DOI: 10.18653/v1/2020.acl-main.704. URL: https://www.aclweb.org/anthology/2020.acl-main.704 (cited on page 22).

Rico Sennrich and Biao Zhang (2019). "Revisiting Low-Resource Neural Machine Translation: A Case Study". *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pages 211–221. DOI: 10.18653/v1/P19-1021. URL: https://www.aclweb.org/anthology/P19-1021 (cited on pages 4, 98).

Rico Sennrich, Barry Haddow, and Alexandra Birch (2016a). "Controlling Politeness in Neural Machine Translation via Side Constraints". *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, pages 35–40. DOI: 10.18653/v1/N16-1005. URL: https://www.aclweb.org/anthology/N16-1005 (cited on pages 28, 31, 33).

Rico Sennrich, Barry Haddow, and Alexandra Birch (2016b). "Improving Neural Machine Translation Models with Monolingual Data". *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, pages 86–96. DOI: 10.18653/v1/P16-1009. URL: https://www.aclweb.org/anthology/P16-1009 (cited on pages 32, 45, 51, 83).

Rico Sennrich, Barry Haddow, and Alexandra Birch (2016c). "Neural Machine Translation of Rare Words with Subword Units". *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, pages 1715–1725.

DOI: 10.18653/v1/P16-1162. URL: https://www.aclweb.org/anthology/P16-1162 (cited on pages 19, 20, 30, 61, 81).

Kashif Shah, Loïc Barrault, and Holger Schwenk (2010). "Translation Model Adaptation by Resampling". *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*. Uppsala, Sweden: Association for Computational Linguistics, pages 392–399. URL: https://www.aclweb.org/anthology/W10-1759 (cited on page 27).

C. E. Shannon (1948). "A mathematical theory of communication". *The Bell System Technical Journal* 27.3, pages 379–423. DOI: 10.1002/j.1538-7305.1948.tb01338.x (cited on page 13).

Steven Shearing, Christo Kirov, Huda Khayrallah, and David Yarowsky (2018). "Improving Low Resource Machine Translation using Morphological Glosses (Non-archival Extended Abstract)". *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*. Boston, MA: Association for Machine Translation in the Americas, pages 132–139. URL: https://www.aclweb.org/anthology/W18-1813 (cited on page 6).

Shiqi Shen, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu (2016). "Minimum Risk Training for Neural Machine Translation". *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics,

pages 1683–1692. DOI: 10.18653/v1/P16-1159. URL: https://www.aclweb.org/anthology/P16-1159 (cited on page 51).

Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi (2018). "RUSE: Regressor Using Sentence Embeddings for Automatic Machine Translation Evaluation". *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*. Belgium, Brussels: Association for Computational Linguistics, pages 751–758. DOI: 10.18653/v1/W18-6456. URL: https://www.aclweb.org/anthology/W18-6456 (cited on page 22).

Raivis Skadiņš, Jörg Tiedemann, Roberts Rozis, and Daiga Deksne (2014). "Billions of Parallel Words for Free: Building and Using the EU Bookshop Corpus". *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), pages 1850–1855. URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/846_Paper.pdf (cited on pages 34, 84).

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov (2014). "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". *Journal of Machine Learning Research* 15.56, pages 1929–1958. URL: http://jmlr.org/papers/v15/srivastava14a.html (cited on page 17).

Felix Stahlberg, Eva Hasler, Aurelien Waite, and Bill Byrne (2016). "Syntactically Guided Neural Machine Translation". *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Berlin,

Germany: Association for Computational Linguistics, pages 299–305. DOI: 10.18653/v1/P16-2049. URL: https://www.aclweb.org/anthology/P16-2049 (cited on page 21).

Felix Stahlberg, Adrià de Gispert, Eva Hasler, and Bill Byrne (2017). "Neural Machine Translation by Minimising the Bayes-risk with Respect to Syntactic Translation Lattices". *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Valencia, Spain: Association for Computational Linguistics, pages 362–368. URL: https://www.aclweb.org/anthology/E17-2058 (cited on page 21).

Felix Stahlberg, James Cross, and Veselin Stoyanov (2018). "Simple Fusion: Return of the Language Model". *Proceedings of the Third Conference on Machine Translation: Research Papers*. Brussels, Belgium: Association for Computational Linguistics, pages 204–211. DOI: 10.18653/v1/W18-6321. URL: https://www.aclweb.org/anthology/W18-6321 (cited on page 52).

David Stallard, Jacob Devlin, Michael Kayser, Yoong Keok Lee, and Regina Barzilay (2012). "Unsupervised Morphology Rivals Supervised Morphology for Arabic MT". *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Jeju Island, Korea: Association for Computational Linguistics, pages 322–327. URL: https://www.aclweb.org/anthology/P12-2063 (cited on page 18).

Miloš Stanojević and Khalil Sima'an (2014). "BEER: BEtter Evaluation as Ranking". *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Baltimore, Maryland, USA: Association for Computational Linguistics, pages 414–419. DOI: 10.3115/v1/W14-3354. URL: https://www.aclweb.org/anthology/W14-3354 (cited on page 22).

Ilya Sutskever, Oriol Vinyals, and Quoc V Le (2014). "Sequence to Sequence Learning with Neural Networks". *Advances in Neural Information Processing Systems*. Edited by Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger. Volume 27. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2014/file/a14ac55a4f27472c5d894ec1c3c743d2-Paper.pdf (cited on page 15).

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna (2016). "Rethinking the Inception Architecture for Computer Vision". *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826. DOI: 10.1109/CVPR.2016.308 (cited on pages 18, 52, 70).

Wolfgang Täger (2011). "The Sentence-Aligned European Patent Corpus". *Proceedings of the 15th International Conference of the European Association for Machine Translation (EAMT)*. Edited by Mikel L. Forcada, Heidi Depraetere, and Vincent Vandeghinste. Leuven, Belgium, pages 177–184. URL: http://mt-archive.info/EAMT-2011-Tager.pdf (cited on page 34).

Kaveh Taghipour, Shahram Khadivi, and Jia Xu (2011). "Parallel Corpus Refinement as an Outlier Detection Algorithm". *Proceedings of the 13th Machine Translation*

Summit (MT Summit XIII). Xiamen, China: International Association for Machine Translation, pages 414–421. URL: http://www.mt-archive.info/MTS-2011-Taghipour.pdf (cited on page 91).

Brian Thompson and Philipp Koehn (2019). "Vecalign: Improved Sentence Alignment in Linear Time and Space". *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pages 1342–1348. DOI: 10.18653/v1/D19-1136. URL: https://www.aclweb.org/anthology/D19-1136 (cited on page 24).

Brian Thompson and Matt Post (2020a). "Automatic Machine Translation Evaluation in Many Languages via Zero-Shot Paraphrasing". *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pages 90–121. DOI: 10.18653/v1/2020.emnlp-main.8. URL: https://www.aclweb.org/anthology/2020.emnlp-main.8 (cited on pages 22, 53).

Brian Thompson and Matt Post (2020b). "Paraphrase Generation as Zero-Shot Multilingual Translation: Disentangling Semantic Similarity from Lexical and Syntactic Diversity". *Proceedings of the Fifth Conference on Machine Translation*. Online: Association for Computational Linguistics, pages 561–570. URL: https://www.aclweb.org/anthology/2020.wmt-1.67 (cited on page 53).

BIBLIOGRAPHY

Brian Thompson, Huda Khayrallah, Antonios Anastasopoulos, Arya D. McCarthy, Kevin Duh, Rebecca Marvin, Paul McNamee, Jeremy Gwinnup, Tim Anderson, and Philipp Koehn (2018). "Freezing Subnetworks to Analyze Domain Adaptation in Neural Machine Translation". *Proceedings of the Third Conference on Machine Translation: Research Papers*. Brussels, Belgium: Association for Computational Linguistics, pages 124–132. DOI: 10.18653/v1/W18-6313. URL: https://www.aclweb.org/anthology/W18-6313 (cited on pages 6, 26, 29).

Brian Thompson, Rebecca Knowles, Xuan Zhang, Huda Khayrallah, Kevin Duh, and Philipp Koehn (2019a). "HABLex: Human Annotated Bilingual Lexicons for Experiments in Machine Translation". *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pages 1382–1387. DOI: 10.18653/v1/D19-1142. URL: https://www.aclweb.org/anthology/D19-1142 (cited on page 6).

Brian Thompson, Jeremy Gwinnup, Huda Khayrallah, Kevin Duh, and Philipp Koehn (2019b). "Overcoming Catastrophic Forgetting During Domain Adaptation of Neural Machine Translation". *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pages 2062–2068. DOI: 10.18653/v1/N19-

1209. URL: https://www.aclweb.org/anthology/N19-1209 (cited on pages 7, 27, 29, 55, 73).

Jörg Tiedemann (2009). "News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces". *Recent Advances in Natural Language Processing*. Edited by N. Nicolov, K. Bontcheva, G. Angelova, and R. Mitkov. Volume V. Borovets, Bulgaria: John Benjamins, Amsterdam/Philadelphia. ISBN: 978 90 272 4825 1 (cited on pages 60, 84).

Jörg Tiedemann (2012). "Parallel Data, Tools and Interfaces in OPUS". *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA), pages 2214–2218. URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf (cited on pages 34, 43, 60, 84).

Jörg Tiedemann and Yves Scherrer (2019). "Measuring Semantic Abstraction of Multilingual NMT with Paraphrase Recognition and Generation Tasks". *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*. Minneapolis, USA: Association for Computational Linguistics, pages 35–42. DOI: 10.18653/v1/W19-2005. URL: https://www.aclweb.org/anthology/W19-2005 (cited on page 22).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). "Attention is All you Need". *Advances in Neural Information Processing Systems 30*. Edited by I. Guyon, U. V.

Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Curran Associates, Inc., pages 5998–6008. URL: http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf (cited on pages 15, 29, 42, 96).

Ashish Venugopal, Jakob Uszkoreit, David Talbot, Franz Och, and Juri Ganitkevitch (2011). "Watermarking the Outputs of Structured Prediction with an application in Statistical Machine Translation." *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Edinburgh, Scotland, UK.: Association for Computational Linguistics, pages 1363–1372. URL: https://www.aclweb.org/anthology/D11-1126 (cited on page 91).

Sami Virpioja, Jaakko J Väyrynen, Mathias Creutz, and Markus Sadeniemi (2007). "Morphology-aware statistical machine translation based on morphs induced in an unsupervised manner". *Proceedings of the Machine Translation Summit XI*. Copenhagen, Denmark, pages 491–498 (cited on page 18).

Andrew Viterbi (1967). "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm". *IEEE Transactions on Information Theory* 13.2, pages 260–269. DOI: 10.1109/TIT.1967.1054010 (cited on page 20).

Rui Wang, Hai Zhao, Bao-Liang Lu, Masao Utiyama, and Eiichiro Sumita (2016). "Connecting Phrase based Statistical Machine Translation Adaptation". *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. Osaka, Japan: The COLING 2016 Organizing Committee,

pages 3135–3145. URL: https://www.aclweb.org/anthology/C16-1295 (cited on page 27).

Rui Wang, Masao Utiyama, Lemao Liu, Kehai Chen, and Eiichiro Sumita (2017). "Instance Weighting for Neural Machine Translation Domain Adaptation". *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, pages 1482–1488. DOI: 10.18653/v1/D17-1155. URL: https://www.aclweb.org/anthology/D17-1155 (cited on page 28).

Wei Wang, Taro Watanabe, Macduff Hughes, Tetsuji Nakagawa, and Ciprian Chelba (2018a). "Denoising Neural Machine Translation Training with Trusted Data and Online Data Selection". *Proceedings of the Third Conference on Machine Translation: Research Papers*. Brussels, Belgium: Association for Computational Linguistics, pages 133–143. DOI: 10.18653/v1/W18-6314. URL: https://www.aclweb.org/anthology/W18-6314 (cited on page 101).

Xinyi Wang, Hieu Pham, Zihang Dai, and Graham Neubig (2018b). "SwitchOut: an Efficient Data Augmentation Algorithm for Neural Machine Translation". *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pages 856–861. DOI: 10.18653/v1/D18-1100. URL: https://www.aclweb.org/anthology/D18-1100 (cited on pages 33, 52).

Xinyi Wang, Hieu Pham, Paul Michel, Antonios Anastasopoulos, Jaime Carbonell, and Graham Neubig (2020). "Optimizing Data Usage via Differentiable Rewards". *International Conference on Machine Learning (ICML)*. URL: https://arxiv.org/abs/1911.10088 (cited on page 101).

Marlies van der Wees, Arianna Bisazza, and Christof Monz (2017). "Dynamic Data Selection for Neural Machine Translation". *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, pages 1400–1410. DOI: 10.18653/v1/D17-1147. URL: https://www.aclweb.org/anthology/D17-1147 (cited on pages 92, 101).

John Wieting and Kevin Gimpel (2018). "ParaNMT-50M: Pushing the Limits of Paraphrastic Sentence Embeddings with Millions of Machine Translations". *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, pages 451–462. DOI: 10.18653/v1/P18-1042. URL: https://www.aclweb.org/anthology/P18-1042 (cited on page 53).

John Wieting, Jonathan Mallinson, and Kevin Gimpel (2017). "Learning Paraphrastic Sentence Embeddings from Back-Translated Bitext". *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, pages 274–285. DOI: 10.18653/v1/D17-1026. URL: https://www.aclweb.org/anthology/D17-1026 (cited on page 53).

John Wieting, Taylor Berg-Kirkpatrick, Kevin Gimpel, and Graham Neubig (2019a). "Beyond BLEU:Training Neural Machine Translation with Semantic Similarity". *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pages 4344–4355. DOI: 10.18653/v1/P19-1427. URL: https://www.aclweb.org/anthology/P19-1427 (cited on page 51).

John Wieting, Kevin Gimpel, Graham Neubig, and Taylor Berg-Kirkpatrick (2019b). "Simple and Effective Paraphrastic Similarity from Parallel Translations". *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pages 4602–4608. DOI: 10.18653/v1/P19-1453. URL: https://www.aclweb.org/anthology/P19-1453 (cited on page 53).

Ronald J. Williams and David Zipser (1989). "A Learning Algorithm for Continually Running Fully Recurrent Neural Networks". *Neural Computation* 1.2, pages 270–280. DOI: 10.1162/neco.1989.1.2.270. eprint: https://doi.org/10.1162/neco.1989.1.2.270. URL: https://doi.org/10.1162/neco.1989.1.2.270 (cited on pages 16, 39).

Joern Wuebker, Spence Green, John DeNero, Saša Hasan, and Minh-Thang Luong (2016). "Models and Inference for Prefix-Constrained Machine Translation". *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for

Computational Linguistics, pages 66–75. DOI: 10.18653/v1/P16-1007. URL: https://www.aclweb.org/anthology/P16-1007 (cited on page 2).

Patrick Xia, Shijie Wu, and Benjamin Van Durme (2020). "Which *BERT? A Survey Organizing Contextualized Encoders". *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pages 7516–7533. DOI: 10.18653/v1/2020.emnlp-main.608. URL: https://www.aclweb.org/anthology/2020.emnlp-main.608 (cited on page 31).

Hainan Xu and Philipp Koehn (2017). "Zipporah: a Fast and Scalable Data Cleaning System for Noisy Web-Crawled Parallel Corpora". *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, pages 2945–2950. DOI: 10.18653/v1/D17-1319. URL: https://www.aclweb.org/anthology/D17-1319 (cited on page 91).

Runxin Xu, Zhuo Zhi, Jun Cao, Mingxuan Wang, and Lei Li (2020). "Volctrans Parallel Corpus Filtering System for WMT 2020". *Proceedings of the Fifth Conference on Machine Translation*. Online: Association for Computational Linguistics, pages 985–990. URL: https://www.aclweb.org/anthology/2020.wmt-1.112 (cited on page 95).

Elizaveta Yankovskaya, Andre Tättar, and Mark Fishel (2019). "Quality Estimation and Translation Metrics via Pre-trained Word and Sentence Embeddings".

Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2). Florence, Italy: Association for Computational Linguistics, pages 101–105. DOI: 10.18653/v1/W19-5410. URL: https://www.aclweb.org/anthology/W19-5410 (cited on page 22).

Dong Yu, Kaisheng Yao, Hao Su, Gang Li, and Frank Seide (2013). "KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition". *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7893–7897. DOI: 10.1109/ICASSP.2013.6639201 (cited on page 70).

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi (2020). "BERTScore: Evaluating Text Generation with BERT". *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=SkeHuCVFDr (cited on page 22).

Chunting Zhou, Xuezhe Ma, Junjie Hu, and Graham Neubig (2019). "Handling Syntactic Divergence in Low-resource Machine Translation". *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pages 1388–1394. DOI: 10.18653/v1/D19-1143. URL: https://www.aclweb.org/anthology/D19-1143 (cited on page 51).

BIBLIOGRAPHY

Xinpeng Zhou, Hailong Cao, and Tiejun Zhao (2015). "Domain Adaptation for SMT Using Sentence Weight". *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*. Edited by Maosong Sun, Zhiyuan Liu, Min Zhang, and Yang Liu. Cham: Springer International Publishing, pages 153–163. ISBN: 978-3-319-25816-4 (cited on page 27).

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen (2016). "The United Nations Parallel Corpus v1.0". *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia: European Language Resources Association (ELRA), pages 3530–3534. URL: https://www.aclweb.org/anthology/L16-1561 (cited on page 34).

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight (2016). "Transfer Learning for Low-Resource Neural Machine Translation". *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, pages 1568–1575. DOI: 10.18653/v1/D16-1163. URL: https://www.aclweb.org/anthology/D16-1163 (cited on pages 29, 30).

Vita

Huda Khayrallah holds a B.A. in Computer Science from the University of California, Berkeley and an M.S.E. in Computer Science from Johns Hopkins University. She joined the Center for Language and Speech Processing and the Computer Science department at Johns Hopkins in the Fall of 2015.